October 5, 2007

DRAFT Standard for Floating-Point Arithmetic P754

Draft 1.5.0

Last modified at 01:29 BST on October 5, 2007.

Sponsor:
Microprocessor Standards Committee

Abstract: This standard specifies interchange and non-interchange formats and methods for binary and decimal floating-point arithmetic in computer programming environments. Exception conditions are defined and default handling of these conditions is specified.

It is intended that an implementation of a floating-point system conforming to this standard can be realized entirely in software, entirely in hardware, or in any combination of software and hardware. For operations specified in the normative part of this standard, numerical results and exceptions are uniquely determined by the values of the input data, sequence of operations, and destination formats, all under user control.

Keywords: computer, floating-point, arithmetic, rounding, format, interchange, number, binary, decimal, subnormal, NaN, significand, exponent.

IEEE Standards Activities Department
Manager, Standards Intellectual Property
445 Hoes Lane

Piscataway, NJ 08854, USA

Patent statement

Attention is called to the possibility that implementation of this standard might require use of subject matter covered by patent rights. By publication of this standard, no position is taken with respect to the existence or validity of any patent rights in connection therewith. The IEEE shall not be responsible for identifying patents or patent applications for which a license might be required to implement an IEEE standard or for conducting inquiries into the legal validity or scope of those patents that are brought to its attention. A patent holder or patent applicant has filed a statement of assurance that it will grant licenses under these rights without compensation or under reasonable rates and nondiscriminatory, reasonable terms and conditions to applicants desiring to obtain such licenses. The IEEE makes no representation as to the reasonableness of rates, terms, and conditions of the license agreements offered by patent holders or patent applicants. Further information may be obtained from the IEEE Standards Department.

Introduction

[This introduction is not a part of DRAFT Standard for Floating-Point Arithmetic P754.]

This standard is a product of the Floating-Point Working Group of, and sponsored by, the Microprocessor Standards Subcommittee of the IEEE Computer Society.

This standard provides a discipline for performing floating-point computation that yields results independent of whether the processing is done in hardware, software, or a combination of the two. For operations specified in the normative part of this standard, numerical results and exceptions are uniquely determined by the values of the input data, the operation, and the destination, all under user control.

This standard defines a family of commercially feasible ways for systems to perform binary and decimal floating-point arithmetic. Among the desiderata that guided the formulation of this standard were:

a) Facilitate movement of existing programs from diverse computers to those that adhere to this standard as well as among those that adhere to this standard.

b) Enhance the capabilities and safety available to users and programmers who, though not expert in numerical methods, might well be attempting to produce numerically sophisticated programs.

c) Encourage experts to develop and distribute robust and efficient numerical programs that are portable, by way of minor editing and recompilation, onto any computer that conforms to this standard and possesses adequate capacity. Together with language controls it should be possible to write programs that produce identical results on all conforming systems.

d) Provide direct support for
   - execution-time diagnosis of anomalies
   - smoother handling of exceptions
   - interval arithmetic at a reasonable cost.

e) Provide for development of
   - standard elementary functions such as *exp* and *cos*
   - high precision (multiword) arithmetic
   - coupled numerical and symbolic algebraic computation.

f) Enable rather than preclude further refinements and extensions.

In programming environments, this standard is also intended to form the basis for a dialog between the numerical community and programming language designers. It is hoped that language-defined methods for the control of expression evaluation and exceptions might be defined in coming years, so that it will be possible to write programs that produce identical results on all conforming systems. However, it is recognized that utility and safety in languages are sometimes antagonists, as are efficiency and portability.

Therefore, it is hoped that language designers will look on the full set of operation, precision, and exception controls described here as a guide to providing the programmer with the ability to portably control expressions and exceptions. It is also hoped that designers will be guided by this standard to provide extensions in a completely portable way.

Participants

The following people participated in the development of this standard:

Dan Zuras, Chair

Aiken, Alex
Applegate, Matthew
Bailey, David
Bass, Steve
Bhandarkar, Dileep
Bhat, Mahesh
Bindel, David
Boldo, Sylvie
Canon, Stephen
Carlough, Steven
Cornea, Marius
Cowlishaw, Mike
Crawford, John
Darcy, Joseph D
Das Sarma, Debjit
Daumas, Marc
Davis, Bob
Davis, Mark
Delp, Dick
Demmel, Jim
Erle, Mark
Fahmy, Hossam
Fasano, J.P.
Fateman, Richard
Feng, Eric
Ferguson, Warren
Fit-Florea, Alex
Fournier, Laurent
Freitag, Chip
Godard, Ivan
Golliver, Roger
Gustafson, David
Hack, Michel
Harrison, John
Hauser, John
Hida, Yozo
Hinds, Chris
Hoare, Graydon
Hough, David
Huck, Jerry
Hull, Jim
Ingrassia, Michael
James, David V
James, Rick
Kahan, William
Kapernick, John
Karpinski, Richard
Kidder, Jeff
Koev, Plamen
Li, Ren-Cang
Liu, Zhishun Alex
Mak, Raymond
Markstein, Peter
Matula, David
Melquiond, Guillaume
Mori, Nobuyoshi
Morin, Ricardo
Nedialkov, Ned
Nelson, Craig
Oberman, Stuart
Okada, Jon
Ollmann, Ian
Parks, Michael
Pittman, Tom
Postpischil, Eric
Riedy, Jason
Schwarz, Eric
Scott, David
Senzig, Don
Sharapov, Ilya
Shearer, Jim
Siu, Michael
Smith, Ron
Stevens, Chuck
Tang, Peter
Taylor, Pamela
Thomas, Jim
Thompson, Brandon
Thrash, Wendy
Toda, Neil
Trong, Son Dao
Tsai, Leonard
Tsen, Charles
Tydeman, Fred
Wang, Liang Kai
Westbrook, Scott
Winkler, Steve
Wood, Anthony
Yalcinalp, Umit
Zemke, Fred
Zimmermann, Paul
Zuras, Dan

The following members of the balloting committee voted on this standard. Balloters might have voted for approval, disapproval, or abstention.

To be supplied by IEEE etc. etc.

Table of contents

List of figures

List of tables

DRAFT Standard for Floating-Point Arithmetic P754

1. Overview

1.1 Scope

This standard specifies formats and methods for floating-point arithmetic in computer systems: standard and extended functions with single, double, extended, and extendable precision, and recommends formats for data interchange. Exception conditions are defined and standard handling of these conditions is specified.

1.2 Inclusions

This standard specifies:

- Formats for binary and decimal floating-point data, for computation and data interchange
- Addition, subtraction, multiplication, division, fused multiply add, square root, compare, and other operations
- Conversions between integer and floating-point formats
- Conversions between different floating-point formats
- Conversions between floating-point formats and external representations as character sequences
- Floating-point exceptions and their handling, including data that are not numbers (NaNs).

1.3 Exclusions

This standard does not specify:

- Formats of integers
- Interpretation of the sign and significand fields of NaNs.

1.4 Purpose

This standard provides a method for computation with floating-point numbers that will yield the same result whether the processing is done in hardware, software, or a combination of the two. The results of the computation will be identical, independent of implementation, given the same input data. Errors, and error conditions, in the mathematical processing will be reported in a consistent manner regardless of implementation.

1.5 Programming environment considerations

This standard specifies floating-point arithmetic in two radices, 2 and 10. A programming environment may conform to this standard in one radix or in both.

This standard does not define all aspects of a conforming programming environment. Such behavior should be defined by a programming language definition supporting this standard, if available, and otherwise by a particular implementation. Some programming language specifications might permit some behaviors to be defined by the implementation.

Language-defined behavior should be defined by a programming language standard supporting this standard. Then all implementations conforming both to this floating-point standard and to that language

standard behave identically with respect to such language-defined behaviors. Standards for languages intended to reproduce results exactly on all platforms are expected to specify behavior more tightly than do standards for languages intended to maximize performance on every platform.

Because this standard requires facilities that are not currently available in common programming languages, the standards for such languages might not be able to fully conform to this standard if they are no longer being revised. If the language can be extended by a function library or class or package to provide a conforming environment, then that extension should define all the language-defined behaviors that would normally be defined by a language standard.

Implementation-defined behavior is defined by a specific implementation of a specific programming environment conforming to this standard. Implementations define behaviors not specified by this standard nor by any relevant programming language standard or programming language extension.

Conformance to this standard is a property of a specific implementation of a specific programming environment, rather than of a language specification.

However a language standard could also be said to conform to this standard if it were constructed so that every conforming implementation of that language also conformed automatically to this standard.

2. Terms and definitions

2.1 Conformance levels

Several keywords are used to differentiate between different levels of requirements and optionality, as follows.

2.1.1 expected: Describes the behavior of the hardware or software in the design models assumed by this specification. Other hardware and software design models may also be implemented.

2.1.2 may: Indicates a course of action permissible within the limits of the standard with no implied preference ("may" means "is permitted to").

2.1.3 shall: Indicates mandatory requirements strictly to be followed in order to conform to the standard and from which no deviation is permitted ("shall" means "is required to").

2.1.4 should: Indicates that among several possibilities, one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required; or that (in the negative form) a certain course of action is deprecated but not prohibited ("should" means "is recommended to").

2.2 Glossary of terms

2.2.1 applicable attribute: The value of an attribute governing a particular instance of execution of a computational operation of this standard. Languages specify how the applicable attribute is determined.

2.2.2 attribute: An implicit parameter to operations of this standard, which a program might statically set in a programming language by specifying a constant value. The term attribute might refer to the parameter (as in "rounding-direction attribute") or its value (as in "roundTowardZero attribute").

2.2.3 basic format: One of the five sets of floating-point representations, three binary and two decimal, whose encodings are specified by this standard, and which are available for arithmetic.

2.2.4 biased exponent: The sum of the exponent and a constant (bias) chosen to make the biased exponent's range nonnegative.

2.2.5 binary floating-point number: A floating-point number with radix two.

2.2.6 block: A language-defined syntactic unit for which a programmer can specify attributes. Language standards might provide means for programs to specify attributes for blocks of varying scopes, even as large as an entire program and as small as a single operation.

2.2.7 canonical encoding: The preferred encoding of a floating-point representation in a format. Applied to declets, significands of finite numbers, infinities, and NaNs, especially in decimal formats.

2.2.8 cohort: In a given format, the set of representations of floating-point numbers with the same numerical value. +0 and −0 are in separate cohorts.

2.2.9 computational operation: An operation that can produce a floating-point result or signal a floating-point exception. Comparisons are computational operations.

2.2.10 correct rounding: This standard's method of converting an infinitely precise result to a floating-point number, as determined by the applicable rounding-direction. A floating-point number so obtained is said to be correctly rounded.

2.2.11 decimal floating-point number: A floating-point number with radix ten.

2.2.12 declet: An encoding of three decimal digits into ten bits using the densely-packed decimal encoding scheme. Of the 1024 possible declets, 1000 canonical declets are produced by computational operations, while 24 non-canonical declets are not produced by computational operations, but are accepted in operands.

2.2.13 denormalized number: See subnormal number.

2.2.14 destination: The location for the result of an operation upon one or more operands. A destination might be either explicitly designated by the user or implicitly supplied by the system (for example, intermediate results in subexpressions or arguments for procedures). Some languages place the results of intermediate calculations in destinations beyond the user's control; nonetheless, this standard defines the result of an operation in terms of that destination's format and the operands' values.

2.2.15 dynamic mode: An optional method of dynamically setting attributes by means of operations of this standard to set, test, save, and restore them.

2.2.16 exception: An event that occurs when an operation has no outcome suitable for every reasonable application. That operation might signal one or more exceptions by invoking the default or, if explicitly requested, a language-defined alternate handling. Note that "event", "exception", and "signal" are defined in diverse ways in different programming environments.

2.2.17 exponent: The component of a finite floating-point representation that signifies the integer power to which the radix is raised in determining the value of that floating-point representation. The exponent e is used when the significand is regarded as an integer digit and fraction field, and the exponent q is used when the significand is regarded as an integer; $e=q+p-1$ where p is the significand length in digits.

2.2.18 extendable precision format: A format with a precision and range that is defined under program control.

2.2.19 extended precision format: A format that extends a supported basic format with wider precision and range and is language-defined or implementation-defined.

2.2.20 external character sequence: A representation of a floating-point datum as a sequence of characters, including the character sequences in floating-point literals in program text.

2.2.21 flag: See status flag.

2.2.22 floating-point datum: A floating-point number or non-number (NaN) that is representable in a floating-point format. In this standard, a floating-point datum is not always distinguished from its representation or encoding.

2.2.23 floating-point number: A finite or infinite number that is representable in a floating-point format. A floating-point datum that is not a NaN. All floating-point numbers, including zeros and infinities, are signed.

2.2.24 floating-point representation: An unencoded member of a floating-point format, representing a finite number, a signed infinity, or a quiet or signaling NaN. A representation of a finite number has three components: a sign, an exponent, and a significand; its numerical value is the signed product of its significand and its radix raised to the power of its exponent.

2.2.25 format: A set of representations of numerical values and symbols, perhaps accompanied by an encoding.

2.2.26 fusedMultiplyAdd: The operation fusedMultiplyAdd(x, y, z) computes $(x\times y)+z$ as if with unbounded range and precision, rounding only once to the destination format.

2.2.27 generic operation: An operation that can take operands of various formats, for which the formats of the results might depend on the formats of the operands.

2.2.28 homogeneous operation: An operation of this standard that takes operands and returns results all in the same format.

2.2.29 implementation-defined: Behavior defined by a specific implementation of a specific programming environment conforming to this standard.

2.2.30 interchange format: A format which has an encoding defined in this standard.

2.2.31 language-defined: Behavior defined by a programming language standard supporting this standard.

2.2.32 NaN: not a number—a symbolic floating-point datum. There are two types of NaN representations: quiet and signaling. Most operations propagate **quiet NaN**s without signaling exceptions, and signal the invalid exception when given a **signaling NaN** operand.

2.2.33 narrower/wider format: If the set of floating-point numbers of one format is a proper subset of another format, the first is called narrower and the second wider. The wider format might have greater precision, range, or (usually) both.

2.2.34 non-computational operation: An operation that neither produces a floating-point result nor signals a floating-point exception..

2.2.35 non-interchange format: A format which does not have an encoding defined in this standard.

2.2.36 normal number: For a particular format, a finite non-zero floating-point number with magnitude greater than or equal to a minimum b^{emin} value. Normal numbers can use the full precision available in a format. In this standard, zero is neither normal nor subnormal.

2.2.37 not a number: See **NaN**.

2.2.38 payload: The diagnostic information contained in a NaN, encoded in part of its trailing significand field.

2.2.39 precision: The number of digits that can be represented in a format, or the number of digits to which a result is rounded.

2.2.40 preferred exponent: For the result of a decimal operation, the value of the exponent q which best preserves the quantum of the operands when the result is exact.

2.2.41 quantum: The quantum of a finite floating-point representation is the value of a unit in the last position of its significand. This is equal to the radix raised to the exponent q.

2.2.42 quiet operation: An operation that never signals any floating-point exception.

2.2.43 radix: The base for the representation of binary or decimal floating-point numbers, two or ten.

2.2.44 result: The floating-point representation or encoding that is delivered to the destination.

2.2.45 signal: When an operation has no outcome suitable for every reasonable application, that operation might signal one or more exceptions by invoking the default handling or, if explicitly requested, a language-defined alternate handling.

2.2.46 significand: A component of a finite floating-point number containing its significant digits. The significand can be thought of as an integer, a fraction, or some other fixed-point form, by choosing an appropriate exponent offset.

2.2.47 status flag: A variable that might take two states, raised or lowered. When raised, a status flag might convey additional system-dependent information, possibly inaccessible to some users. The operations of this standard, when exceptional, can as a side effect raise some of the following status flags: inexact, underflow, overflow, divideByZero, and invalid.

2.2.48 storage format: One of the two sets of floating-point representations, one binary and one decimal, whose encodings are specified by the standard, and which might not be available for arithmetic.

2.2.49 subnormal number: In a particular format, a non-zero floating-point number with magnitude less than the magnitude of that format's smallest normal number. A subnormal number does not use the full precision available to normal numbers of the same format.

2.2.50 supported format: A format provided in the programming environment and implemented in conformance with the requirements of this standard. Thus, a programming environment might provide more formats than it supports, as only those implemented in accordance with the standard are said to be supported. An integer format is said to be supported if conversions between that format and supported floating-point formats are provided in conformance with this standard.

2.2.51 trailing significand: A component of an encoded binary or decimal floating-point format containing all the significand digits except the leading digit. In these formats, the biased exponent or combination field encodes the leading significand digit.

2.2.52 user: Any person, hardware, or program not itself specified by this standard, having access to and controlling those operations of the programming environment specified in this standard.

2.2.53 widenTo method: A method used by a programming language to determine the formats for evaluating generic operators and functions. Some **widenTo** methods take advantage of the extra range and precision of wide formats without requiring the program to be written with explicit conversions.

2.2.54 width of an operation: The format of the destination of an operation specified by this standard; it will be one of the supported formats provided by an implementation in conformance to this standard.

2.3 Abbreviations and acronyms

LSB	least significant bit
MSB	most significant bit
NaN	not a number
qNaN	quiet NaN
sNaN	signaling NaN

3. Formats

3.1 Overview: formats and conformance

This clause defines several kinds of standard floating-point formats, in two radices, 2 and 10. All the formats specified by this standard are fixed-width. The precision and range of a fixed-width format are determinable from the program text, and the corresponding encoding is usually defined so that all members have the same size in storage.

Formats defined by this standard are interchange or non-interchange:

- **interchange formats** are formats with encodings defined in this standard. They are widely available for storage and for data interchange among platforms. The format names used in this standard are not usually those used in programming environments. Interchange formats defined by this standard are:
  - **basic formats**, which are interchange formats available for arithmetic. This standard defines three basic binary floating-point formats in lengths of 32, 64, and 128 bits, and two basic decimal floating-point formats in lengths of 64 and 128 bits
  - **storage formats**, which are narrow interchange formats. This standard defines one binary storage floating-point format of 16 bits length, and one decimal storage floating-point format of 32 bits length; language standards permitting computation upon storage formats should support such computations in a wider format
  - **formats for extended and extendable precision,** which extend the encodings of the basic and storage formats to support the interchange of floating-point data at additional widths.
- **non-interchange formats** are extended and extendable precision formats whose encodings are not defined in this standard but which are available for arithmetic. None are required by this standard. Where required, interchange of data in these formats should be done using a suitably large interchange format or external character sequences that meet the requirements of 5.12.

A programming environment conforms to this standard, in a particular radix, by implementing one or more of the basic formats of that radix. The choice of which of this standard's formats to support is language-defined or, if the relevant language standard is silent or defers to the implementation, implementation-defined.

A conforming implementation of any format shall:

- provide means to initialize and store that format
- provide conversions between that format and all other supported formats.

A conforming implementation of a format available for arithmetic shall:

- provide all the operations of this standard, as defined in clause 5, for that format.

3.2 Specification levels

Floating-point arithmetic is a systematic approximation of real arithmetic, as illustrated in Table 1. Floating-point arithmetic can only represent a finite subset of the continuum of real numbers. Consequently certain properties of real arithmetic, such as associativity of addition, do not always hold for floating-point arithmetic.

Table 1—Relationships between different specification levels for a particular format

Level 1	**{−∞ … 0 … +∞}**	Extended real numbers.
many-to-one ↓	*rounding*	↑ projection (except for NaN)
Level 2	**{−∞ … −0} ∪ {+0 … +∞} ∪ NaN**	Floating-point data — an algebraically closed system.
one-to-many ↓	*representation specification*	↑ many-to-one
Level 3	**(*sign*, *exponent*, *significand*) ∪ {−∞, +∞} ∪ qNaN ∪ sNaN**	Representations of floating-point data.
one-to-many ↓	*encoding for representations of floating-point data*	↑ many-to-one
Level 4	**0111000…**	Bit strings.

The mathematical structure underpinning the arithmetic in this standard is the extended reals, that is, the set of real numbers together with positive and negative infinity. For a given format, the process of *rounding* (see 4) maps an extended real number to a *floating-point number* included in that format. A *floating-point datum*, which can be a signed zero, finite non-zero number, signed infinity, or not-a-number (NaN), can be mapped to one or more *representations of floating-point data* in a format.

The representations of floating-point data in a format consist of:

— triples (*sign*, *exponent*, *significand*); in radix *b*, the floating-point number represented by a triple is $(-1)^{sign} \times b^{exponent} \times significand$

— +∞, −∞

— qNaN (quiet), sNaN (signaling).

An *encoding* maps a representation of a floating-point datum to a bit string. An encoding might map some representations of floating-point data to more than one bit string. Multiple NaN bit strings should be used to store retrospective diagnostic information (see 6.2).

3.3 Sets of floating-point data

This subclause specifies the sets of floating-point data representable within all floating-point formats; the encodings for representations of floating-point data in interchange formats are discussed in 3.4, 3.5, and 3.7. The set of finite floating-point numbers representable within a particular format is determined by the following integer parameters:

— *b* = the radix, 2 or 10

— *p* = the number of significant digits (precision)

— *emax* = the maximum exponent *e*

— *emin* = the minimum exponent *e*

emin shall be 1 − *emax* for all formats.

The values of these parameters for each basic and storage format are given in Table 2, which refers to each format by the number of bits in its encoding. Constraints on these parameters for extended and extendable precision formats are given in 3.6.

Within each format, the following floating-point data shall be represented:

— Signed zero and non-zero floating-point numbers of the form $(-1)^s \times b^e \times m$, where:
  — s is 0 or 1
  — e is any integer $emin \le e \le emax$
  — m is a number represented by a digit string of the form $d_0 \bullet d_1 d_2 \ldots d_{p-1}$ where d_i is an integer digit $0 \le d_i < b$ (therefore $0 \le m < b$)
— Two infinities, $+\infty$ and $-\infty$
— Two NaNs, qNaN (quiet) and sNaN (signaling).

These are the only floating-point data represented.

In the foregoing description, the significand m is viewed in a scientific form, with the radix point immediately following the first digit. It is also convenient for some purposes to view the significand as an integer; in which case the finite floating-point numbers are described thus:

— Signed zero and non-zero floating-point numbers of the form $(-1)^s \times b^q \times c$, where
  — s is 0 or 1
  — q is any integer $emin \le q + p - 1 \le emax$
  — c is a number represented by a digit string of the form $d_0 d_1 d_2 \ldots d_{p-1}$ where d_i is an integer digit $0 \le d_i < b$ (c is therefore an integer with $0 \le c < b^p$).

This view of the significand as an integer c, with its corresponding exponent q, describes exactly the same set of zero and non-zero floating-point numbers as the view in scientific form. (For finite floating-point numbers, $e = q + p - 1$ and $m = c \times b^{1-p}$.)

The smallest positive *normal* floating-point number is b^{emin} and the largest is $b^{emax} \times (b - b^{1-p})$. The non-zero floating-point numbers for a format with magnitude less than b^{emin} are called *subnormal* because their magnitudes lie between zero and the smallest normal magnitude. Subnormal numbers are distinguished from normal numbers because of reduced precision and, in binary interchange formats, because of different encoding methods. Every finite floating-point number is an integral multiple of the smallest subnormal magnitude $b^{emin} \times b^{1-p}$.

For a floating-point number that has the value zero, the sign bit s provides an extra bit of information. Although all formats have distinct representations for +0 and −0, the sign of a zero is significant in some circumstances, such as division by zero, but not in others (see 6.3). Binary interchange formats have just one representation each for +0 and −0, but decimal formats have many. In this standard, 0 and ∞ are written without a sign when the sign is not important.

Table 2—Parameters defining basic and storage format floating-point numbers

	Binary format (b=2)				**Decimal format** (b=10)		
parameter	**binary16** storage	**binary32** basic	**binary64** basic	**binary128** basic	**decimal32** storage	**decimal64** basic	**decimal128** basic
p, digits	11	24	53	113	7	16	34
emax	+15	+127	+1023	+16383	+96	+384	+6144
emin	−14	−126	−1022	−16382	−95	−383	−6143

3.4 Binary interchange format encodings

Each floating-point number has just one encoding in a binary interchange format. To make the encoding unique, in terms of the parameters in 3.3, the value of the significand m is maximized by decreasing e until either $e = emin$ or $m \geq 1$. After this normalization process is done, if $e = emin$ and $0 < m < 1$, the floating-point number is subnormal. Subnormal numbers (and zero) are encoded with a reserved biased exponent value.

Representations of floating-point data in the binary interchange formats are encoded in k bits in the following three fields ordered as shown in Figure 3.1:

a) 1-bit sign S
b) w-bit biased exponent $E = e + bias$
c) $(t = p - 1)$-bit trailing significand digit string $T = d_1 d_2 \ldots d_{p-1}$; the leading bit of the significand, d_0, is implicitly encoded in the biased exponent E.

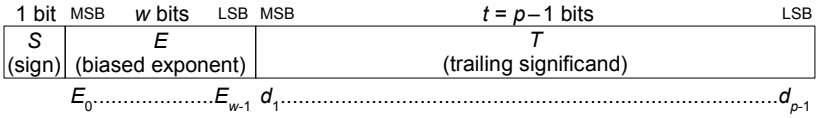

Figure 3.1—Binary interchange floating-point format

The values of k, t, w, and $bias$ for the binary basic and storage formats are listed in Table 3.

The range of the encoding's biased exponent E shall include:

— Every integer between 1 and $2^w - 2$, inclusive, to encode normal numbers
— The reserved value 0 to encode ±0 and subnormal numbers
— The reserved value $2^w - 1$ to encode ±∞ and NaNs.

The representation r of the floating-point datum, and value v of the floating-point datum represented, are inferred from the constituent fields thus:

a) If $E = 2^w - 1$ and $T \neq 0$, then r is qNaN or sNaN and v is NaN regardless of S (see 6.2.1).
b) If $E = 2^w - 1$ and $T = 0$, then r and $v = (-1)^S \times +\infty$.
c) If $1 \leq E \leq 2^w - 2$, then r is $(S, (E - bias), (1 + 2^{1-p} \times T))$;
the value of the corresponding floating-point number is $v = (-1)^S \times 2^{E-bias} \times (1 + 2^{1-p} \times T)$;
thus normal numbers have an implicit leading significand bit of 1.
d) If $E = 0$ and $T \neq 0$, then r is $(S, emin, (0 + 2^{1-p} \times T))$;
the value of the corresponding floating-point number is $v = (-1)^S \times 2^{emin} \times (0 + 2^{1-p} \times T)$;
thus subnormal numbers have an implicit leading significand bit of 0.
e) If $E = 0$ and $T = 0$, then r is $(S, emin, 0)$ and $v = (-1)^S \times +0$ (signed zero, see 6.3).

Table 3—Binary basic and storage format encoding parameters

	Format name			
Parameter (widths in bits)	**binary16**	**binary32**	**binary64**	**binary128**
k, storage width	16	32	64	128
t, trailing significand width	10	23	52	112
w, biased exponent field width	5	8	11	15
$E - e$, bias	15	127	1023	16383

3.5 Decimal interchange format encodings

Unlike in a binary floating-point format, in a decimal floating-point format a number might have multiple representations. The set of representations a floating-point number maps to is called the floating-point number's *cohort*; the members of a cohort are distinct *representations* of the same floating-point number. For example, if c is a multiple of 10 and q is less than its maximum allowed value, then (s, q, c) and (s, $q+1$, $c/10$) are two representations for the same floating-point number and are members of the same cohort.

While numerically equal, different members of a cohort can be distinguished by the decimal-specific operations (see 5.3.2, 5.5.2, and 5.7.3). The cohorts of different floating-point numbers might have different numbers of members. If a finite non-zero number's representation has n decimal digits from its most significant non-zero digit to its least significant non-zero digit, the representation's cohort will have at most $p-n+1$ members where p is the number of digits of precision in the format.

For example, a one-digit floating-point number might have up to p different representations while a p-digit floating-point number with no trailing zeros has only one representation. (An n-digit floating-point number might have fewer than $p-n+1$ members in its cohort if it is near the extremes of the format's exponent range.) A zero has a much larger cohort: the cohort of +0 contains a representation for each exponent, as does the cohort of −0.

For decimal arithmetic, besides specifying a numerical result, the arithmetic operands also select a member of the result's cohort according to 5.2. Decimal applications can make use of the additional information cohorts convey.

Representations of floating-point data in the decimal interchange formats are encoded in k bits in the following three fields, whose detailed layouts are described later.

a) 1-bit sign S.

b) A $w+5$ bit combination field G encoding classification and, if the encoded datum is a finite number, the exponent q and four significand bits (1 or 3 of which are implied). The biased exponent E is a $w+2$ bit quantity $q+bias$, where the value of the first two bits of the biased exponent taken together is either 0, 1, or 2.

c) A t-bit trailing significand field T which contains $J\times 10$ bits and contains the bulk of the significand. When this field is combined with the leading significand bits from the combination field, the format encodes a total of $p = 3\times J+1$ decimal digits.

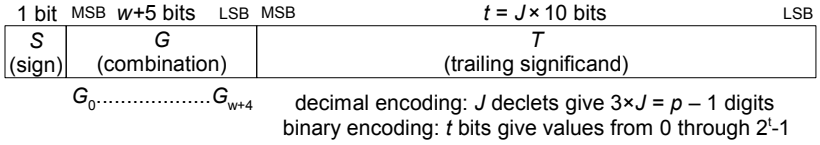

Figure 3.2—Decimal interchange floating-point formats

The values of k, t, w, and $bias$ for the decimal basic and storage formats are listed in Table 4:

Table 4—Decimal basic and storage format encoding parameters

	Format name		
Parameter (widths in bits)	**decimal32**	**decimal64**	**decimal128**
k, storage width	32	64	128
t, trailing significand width	20	50	110
$w+5$, combination field width	11	13	17
$E-q$, bias	101	398	6176

The representation r of the floating-point datum, and value v of the floating-point datum represented, are inferred from the constituent fields, thus:

a) If G_0 through G_4 are 11111, then v is NaN regardless of S. Furthermore, if G_5 is 1, then r is sNaN; otherwise r is qNaN. The remaining bits of G are ignored, and T constitutes the NaN's payload, which can be used to distinguish various NaNs.

   The NaN payload is encoded similarly to finite numbers described below, with G treated as though all bits were zero. The payload corresponds to the significand of finite numbers, interpreted as an integer with a maximum value of $10^{(3\times J)}-1$, and the exponent field is ignored (it is treated as if it were zero). A NaN is in its preferred (canonical) representation if the bits G_6 through G_{w+4} are zero and the encoding of the payload is canonical.

b) If G_0 through G_4 are 11110 then r and $v = (-1)^S \times +\infty$. The values of the remaining bits in G, and T, are ignored. The two canonical representations of infinity have bits G_5 through $G_{w+4} = 0$, and $T = 0$.

c) For finite numbers, r is (S, $E-bias$, C) and $v = (-1)^S \times 10^{(E-bias)} \times C$, where C is the concatenation of the leading significand digit from the combination field G and the trailing significand field T and the biased exponent E is encoded in the combination field. The encoding within these fields depends on whether the significand uses the decimal or the binary encoding.

   1) If the significand uses the *decimal* encoding, then the least significant w bits of the exponent are G_5 through G_{w+4}. The most significant two bits of the biased exponent and the decimal digit string $d_0\, d_1 \ldots d_{p-1}$ of the significand are formed from bits G_0 through G_4 and T as follows:

      i) When the most significant five bits of G are 110xx or 1110x, the leading significand digit d_0 is $8+G_4$, a value 8 or 9, and the leading biased exponent bits are $2G_2+G_3$, a value 0, 1, or 2.

      ii) When the most significant five bits of G are 0xxxx or 10xxx, the leading significand digit d_0 is $4G_2+2G_3+G_4$, a value in the range 0–7, and the leading biased exponent bits are $2G_0+G_1$, a value 0, 1, or 2. Consequently if T is 0 and the most significant five bits of G are 00000, 01000, or 10000, then $v = (-1)^S \times +0$.

      The $p-1 = 3\times J$ decimal digits $d_1 \ldots d_{p-1}$ are encoded by T which contains J declets encoded in densely-packed decimal.

      A canonical significand has only canonical declets, as shown in Tables 5 and 6. Computational operations produce only the 1000 canonical declets, but also accept the 24 non-canonical declets in operands.

   2) Alternatively, if the significand uses the *binary* encoding, then:

      i) If G_0 and G_1 together are one of 00, 01, or 10, then the biased exponent E is formed from G_0 through G_{w+1} and the significand is formed from bits G_{w+2} through the end of the encoding (including T).

      ii) If G_0 and G_1 together are 11 and G_2 and G_3 together are one of 00, 01, or 10, then the biased exponent E is formed from G_2 through G_{w+3} and the significand is formed by prefixing the 4 bits $(8+G_{w+4})$ to T.

      The maximum value of the binary-encoded significand is the same as that of the equivalent decimal-encoded significand; that is, $10^{(3\times J+1)}-1$ (or $10^{(3\times J)}-1$ when T is used as the payload of a NaN). If the value exceeds the maximum, the significand c is non-canonical and the value used for c is zero.

   Computational operations produce only canonical significands, but also accept non-canonical significands in operands.

Decoding densely-packed decimal: Table 5 decodes a declet, with 10 bits $\mathbf{b_{(0)}}$ to $\mathbf{b_{(9)}}$, into 3 decimal digits $\mathbf{d_{(1)}}$, $\mathbf{d_{(2)}}$, $\mathbf{d_{(3)}}$. The first column is in binary and an "x" denotes "don't care". Thus all 1024 possible 10-bit patterns shall be accepted and mapped into 1000 possible 3-digit combinations with some redundancy.

Table 5—Decoding 10-bit densely-packed decimal to 3 decimal digits

$\mathbf{b_{(6)}}$, $\mathbf{b_{(7)}}$, $\mathbf{b_{(8)}}$, $\mathbf{b_{(3)}}$, $\mathbf{b_{(4)}}$	$\mathbf{d_{(1)}}$	$\mathbf{d_{(2)}}$	$\mathbf{d_{(3)}}$
0 x x x x	$4b_{(0)} + 2b_{(1)} + b_{(2)}$	$4b_{(3)} + 2b_{(4)} + b_{(5)}$	$4b_{(7)} + 2b_{(8)} + b_{(9)}$
1 0 0 x x	$4b_{(0)} + 2b_{(1)} + b_{(2)}$	$4b_{(3)} + 2b_{(4)} + b_{(5)}$	$8 + b_{(9)}$
1 0 1 x x	$4b_{(0)} + 2b_{(1)} + b_{(2)}$	$8 + b_{(5)}$	$4b_{(3)} + 2b_{(4)} + b_{(9)}$
1 1 0 x x	$8 + b_{(2)}$	$4b_{(3)} + 2b_{(4)} + b_{(5)}$	$4b_{(0)} + 2b_{(1)} + b_{(9)}$
1 1 1 0 0	$8 + b_{(2)}$	$8 + b_{(5)}$	$4b_{(0)} + 2b_{(1)} + b_{(9)}$
1 1 1 0 1	$8 + b_{(2)}$	$4b_{(0)} + 2b_{(1)} + b_{(5)}$	$8 + b_{(9)}$
1 1 1 1 0	$4b_{(0)} + 2b_{(1)} + b_{(2)}$	$8 + b_{(5)}$	$8 + b_{(9)}$
1 1 1 1 1	$8 + b_{(2)}$	$8 + b_{(5)}$	$8 + b_{(9)}$

Encoding densely-packed decimal: Table 6 encodes 3 decimal digits $\mathbf{d_{(1)}}$, $\mathbf{d_{(2)}}$, and $\mathbf{d_{(3)}}$, each having 4 bits which can be expressed by a second subscript $\mathbf{d_{(1,0:3)}}$, $\mathbf{d_{(2,0:3)}}$, and $\mathbf{d_{(3,0:3)}}$, where bit 0 is the most significant and bit 3 the least significant, into a declet, with 10 bits $\mathbf{b_{(0)}}$ to $\mathbf{b_{(9)}}$. Computational operations generate only the 1000 canonical 10-bit patterns defined by Table 6.

Table 6—Encoding 3 decimal digits to 10-bit densely-packed decimal

$\mathbf{d_{(1,0)}}$, $\mathbf{d_{(2,0)}}$, $\mathbf{d_{(3,0)}}$	$\mathbf{b_{(0)}}$, $\mathbf{b_{(1)}}$, $\mathbf{b_{(2)}}$	$\mathbf{b_{(3)}}$, $\mathbf{b_{(4)}}$, $\mathbf{b_{(5)}}$	$\mathbf{b_{(6)}}$	$\mathbf{b_{(7)}}$, $\mathbf{b_{(8)}}$, $\mathbf{b_{(9)}}$
0 0 0	$d_{(1,1:3)}$	$d_{(2,1:3)}$	0	$d_{(3,1:3)}$
0 0 1	$d_{(1,1:3)}$	$d_{(2,1:3)}$	1	0, 0, $d_{(3,3)}$
0 1 0	$d_{(1,1:3)}$	$d_{(3,1:2)}$, $d_{(2,3)}$	1	0, 1, $d_{(3,3)}$
0 1 1	$d_{(1,1:3)}$	1, 0, $d_{(2,3)}$	1	1, 1, $d_{(3,3)}$
1 0 0	$d_{(3,1:2)}$, $d_{(1,3)}$	$d_{(2,1:3)}$	1	1, 0, $d_{(3,3)}$
1 0 1	$d_{(2,1:2)}$, $d_{(1,3)}$	0, 1, $d_{(2,3)}$	1	1, 1, $d_{(3,3)}$
1 1 0	$d_{(3,1:2)}$, $d_{(1,3)}$	0, 0, $d_{(2,3)}$	1	1, 1, $d_{(3,3)}$
1 1 1	0, 0, $d_{(1,3)}$	1, 1, $d_{(2,3)}$	1	1, 1, $d_{(3,3)}$

The 24 non-canonical patterns of the form 01x11x111x, 10x11x111x, or 11x11x111x (where an "x" denotes "don't care") are not generated in the result of a computational operation. However, as listed in Table 5, these 24 bit patterns do map to values in the range 0−999. The bit pattern in a NaN significand can affect how the NaN is propagated (see 6.2).

3.6 Extended and extendable precisions

Extended and extendable precision formats extend the precisions available for arithmetic beyond those described in 3.4 and 3.5. Specifically:

— an **extended precision format** is a format that extends a supported basic format with wider precision and range and is language-defined or implementation-defined

— an **extendable precision format** is a format with a precision and range that is defined under program control.

These formats are characterized by the parameters *b, p, emax,* and *emin*, which may match those of an interchange format and shall:

— provide all the representations of floating-point data defined in terms of those parameters in 3.2 and 3.3

— provide all the operations of this standard, as defined in clause 5, for that format.

This standard does not require an implementation to provide any extended or extendable precision format. Encodings for storage and arithmetic using these formats are implementation-defined, but should be fixed width and may match those of an interchange format.

Language standards should define mechanisms supporting extendable precision for each supported radix. Language standards supporting extendable precision shall permit programs to specify *p* and *emax* and shall define $emin = 1 - emax$. Language standards should also allow the specification of an extendable precision by specifying *p* alone; in this case *emax* should be defined to be $\geq 1000 \times p$.

Language standards or implementations should support an extended precision format that extends the widest basic format that is supported in that radix. Table 7 specifies the minimum precision and exponent range of the extended precision format for each basic format.

Table 7—Extended format parameters for floating-point numbers

	Extended formats associated with:				
Parameter	**binary32**	**binary64**	**binary128**	**decimal64**	**decimal128**
p digits ≥	32	64	128	20	43
emax ≥	1023	16383	65535	6144	24576
emin ≤	−1022	−16382	−65534	−6143	−24575

NOTE — The minimum exponent range is that of the next wider basic format, if there is one, while the minimum precision is intermediate between the widest supported basic format and the next wider basic format.

3.7 Interchange formats for extended and extendable precision

These formats supplement the interchange formats of 3.4 and 3.5 to support the interchange of floating-point data at additional fixed widths. In each radix, the precision and range of an interchange format is defined by its size; interchange of a floating-point datum of a given size is therefore always exact with no possibility of overflow or underflow.

The encodings for the interchange formats are as described in 3.4 and 3.5, with precision p defined as a function of the format width k in bits, leading to the other parameters as shown in Table 8:

Table 8—Parameters for interchange formats

	Radix	
Parameter	**binary**	**decimal**
k, width in bits	≥ 128; multiple of 32	≥ 32; multiple of 32
p, precision in digits	$k - \text{int}(4 \times \log2(k)) + 13$	$k \times 9 / 32 - 2$
t, trailing significand width	$p - 1$	$(p - 1) \times 10 / 3$
w, exponent field width	$k - t - 1$	$k - t - 6$
emax	$2^{(w-1)} - 1$	$3 \times 2^{(w-1)}$
emin	$1 - emax$	$1 - emax$
bias	$E - e = emax$	$E - q = emax + p - 2$

The function int() in Table 8 is convertToIntegerTiesToEven().

Examples of some specific interchange formats are shown in Table 9:

Table 9—Examples of interchange formats

	Parameter		
Format	k, width in bits	p, precision in digits	*emax*
binary128	128	113	16383
binary256	256	237	262143
binary512	512	489	4194303
binary1024	1024	997	67108863
decimal96	96	25	1536
decimal128	128	34	6144
decimal192	192	52	98304
decimal256	256	70	1572864

4. Attributes and rounding

4.1 Attribute specification

An attribute is logically associated with a program block to modify its numerical and exception semantics. With attribute specification, a user can specify a constant value for an attribute parameter.

Some attributes have the effect of an implicit parameter to most individual operations of this standard; language standards shall provide support for:

— rounding-direction attributes (see 4.3)

and should provide support for:

— alternate exception handling attributes (see 8).

Other attributes change the mapping of language expressions into operations of this standard; language standards that permit more than one such mapping should provide support for:

— widenTo attributes (see 10.3)
— value-changing optimization attributes (see 10.4)
— reproducibility attributes (see 11).

For attribute specification, the implementation shall provide language-defined means, such as compiler directives, to specify a constant value for the attribute parameter for all standard operations in a block; the scope of the attribute value is the block with which it is associated. Language standards shall provide for constant specification of the default and each specific value of the attribute.

4.2 Dynamic modes for attributes

Attributes in this standard shall be supported with the constant specification of 4.1. Particularly to support debugging, language standards should also support dynamic-mode specification for some or all attributes.

With dynamic-mode specification, a user can specify that the attribute parameter assumes the value of a dynamic-mode variable whose value might not be known until program execution. This standard does not specify the underlying implementation mechanisms for constant attributes or dynamic modes.

For dynamic-mode specification, the implementation shall provide language-defined means to specify that the attribute parameter assumes the value of a dynamic-mode variable for all standard operations within the scope of the dynamic-mode specification in a block. The implementation initializes a dynamic-mode variable to the default value for the dynamic mode. Within its language-defined (dynamic) scope, changes to the value of a dynamic-mode variable are under the control of the user via the operations in 9.3.1 and 11.

The following aspects of dynamic-mode variables are language-defined; language standards might explicitly defer the definitions to implementations:

— whether the dynamic-mode parameter assumes the default attribute value or the value of a dynamic-mode variable
— precedence of static attribute specifications and dynamic-mode assignments
— the effect of changing the value of the dynamic-mode variable in an asynchronous event, such as in another thread or signal handler
— whether the value of the dynamic-mode variable can be determined by non-programmatic means, such as a debugger.

4.3 Rounding-direction attributes

Rounding takes a number regarded as infinitely precise and, if necessary, modifies it to fit in the destination's format while signaling the inexact exception (see 7.6), underflow, or overflow when appropriate. Every operation shall be performed as if it first produced an intermediate result correct to

infinite precision and with unbounded range, and then rounded that result according to one of the attributes in this clause.

The rounding-direction attribute affects all computational operations that might be inexact. Inexact numeric floating-point results always have the same sign as the unrounded result.

The rounding-direction attribute affects the signs of exact zero sums (see 6.3), and also affects the thresholds beyond which overflow (see 7.4) and underflow (see 7.5) are signaled.

Implementations supporting both decimal and binary formats shall provide separate rounding-direction attributes for binary and decimal, the binary rounding direction and the decimal rounding direction. Operations returning results in a floating-point format shall use the rounding-direction attribute associated with the radix of the results. Operations converting from an operand in a floating-point format to a result in integer format or external character sequence format shall use the rounding-direction attribute associated with the radix of the operand.

4.3.1 Rounding-direction attributes to nearest

In the following two rounding-direction attributes an infinitely precise result with magnitude at least $b^{emax}(b-\frac{1}{2}b^{1-p})$ shall round to ∞ with no change in sign; here *emax* and *p* are determined by the destination format (see 3.3). With:

- roundTiesToEven, the floating-point number nearest to the infinitely precise result shall be delivered; if the two nearest floating-point numbers bracketing an unrepresentable infinitely precise result are equally near, the one with an even least significant digit shall be delivered
- roundTiesToAway, the floating-point number nearest to the infinitely precise result shall be delivered; if the two nearest floating-point numbers bracketing an unrepresentable infinitely precise result are equally near, the one with larger magnitude shall be delivered.

4.3.2 Directed rounding attributes

Three other user-selectable rounding-direction attributes are defined, the directed rounding attributes roundTowardPositive, roundTowardNegative, and roundTowardZero. With:

- roundTowardPositive, the result shall be the format's floating-point number (possibly $+\infty$) closest to and no less than the infinitely precise result
- roundTowardNegative, the result shall be the format's floating-point number (possibly $-\infty$) closest to and no greater than the infinitely precise result
- roundTowardZero, the result shall be the format's floating-point number closest to and no greater in magnitude than the infinitely precise result.

4.3.3 Rounding attribute requirements

An implementation of this standard shall provide roundTiesToEven and the three directed rounding attributes. A decimal implementation of this standard shall provide roundTiesToAway as a user-selectable rounding-direction attribute. The rounding attribute roundTiesToAway is not required for binary.

The roundTiesToEven rounding-direction attribute shall be the default rounding-direction attribute for results in binary formats. The default rounding-direction attribute for results in decimal formats is language-defined, but should be roundTiesToEven.

5. Operations

5.1 Overview

All conforming implementations of this standard shall provide the operations listed in this clause for all supported floating-point formats available for arithmetic. Each of the computational operations that return a numeric result specified by this standard shall be performed as if it first produced an intermediate result correct to infinite precision and with unbounded range, and then rounded that intermediate result to fit in the destination's format (see 4 and 7). Clause 6 augments the following specifications to cover ±0, ±∞, and NaN; clause 7 describes default exception handling.

In this standard, operations are written as named functions; in a specific programming environment they might be represented by operators, or by families of format-specific functions, or by generic functions whose names might differ from those in this standard.

Operations are broadly classified in four groups according to the types of results and exceptions they produce:

- general-computational operations produce floating-point results, round all results according to clause 4, and might signal the floating-point exceptions of clause 7
- quiet-computational operations produce floating-point results and do not signal floating-point exceptions
- signaling-computational operations produce no floating-point results and might signal floating-point exceptions; comparisons are signaling-computational operations
- non-computational operations do not produce floating-point results and do not signal floating-point exceptions.

Operations in the first three groups are referred to collectively as "computational operations".

Operations are also classified in two ways according to the relationship between the result format and the operand formats:

- homogeneous operations, in which the floating-point operands and floating-point result are all of the same format
- *formatOf* operations, which indicate the format of the result, independent of the formats of the operands.

Language standards might permit other kinds of operations and combinations of operations in expressions. By their expression evaluation rules, languages specify when and how such operations and expressions are mapped into the operations of this standard.

In the operation descriptions that follow, operand and result formats are indicated by:

- *source* to represent homogeneous floating-point operand formats
- *source1, source2, source3* to represent non-homogeneous floating-point operand formats
- *int* to represent integer operand formats
- *boolean* to represent a value of 0 or 1 (*false* or *true*)
- *enum* to represent one of a small set of enumerated values
- *logBFormat* to represent a type for the destination of the *logB* operation and the scale exponent operand of the *scaleB* operation
- *integralFormat* to represent the scale factor in scaled products (see 9.4)
- *decimalCharacterSequence* to represent a decimal character sequence
- *hexCharacterSequence* to represent a hexadecimal character sequence
- *conversionSpecification* to represent a language dependent conversion specification
- *decimalType* to represent a supported decimal floating-point type
- *decimalEncodingType* to represent a decimal floating-point type encoded in decimal

- *binaryEncodingType* to represent a decimal floating-point type encoded in binary
- *exceptionGroupType* to represent a set of exceptions
- *flagsType* to represent a set of status flags
- *binaryRoundingDirectionType* to represent the rounding direction for binary
- *decimalRoundingDirectionType* to represent the rounding direction for decimal
- *modeGroupType* to represent dynamically-specifiable modes.

formatOf indicates that the name of the operation specifies the floating-point destination *format*, which might be different from the floating-point operands' formats. There are *formatOf* versions of these operations for every supported floating-point format available for arithmetic.

intFormatOf indicates that the name of the operation specifies the integer destination format.

In the operation descriptions that follow, languages define which of their types correspond to operands and results called *int, intFormatOf, characterSequence,* or *conversionSpecification*. Languages with both signed and unsigned integer types should support both signed and unsigned *int* and *intFormatOf* operands and results.

5.2 Decimal exponent calculation

As discussed in 3.5, a floating-point number might have multiple representations in a decimal format. Therefore, decimal arithmetic involves not only computing the proper numerical result but also selecting the proper member of that floating-point number's cohort.

Except for the quantize operation, the value of a floating-point result (and hence its cohort) is determined by the operation and the operands' values; it is never dependent on the representation or encoding of an operand.

The selection of a particular representation for a floating-point result is dependent on the operands' representations, as described below, but is not affected by their encoding.

For all computational operations except quantize, if the result is inexact the cohort member of least possible exponent is used to get the longest possible significand. If the result is exact, the cohort member is selected based on the preferred exponent for a result of that operation, a function of the exponents of the inputs. Thus for finite x, depending on the representation of zero, $0+x$ might result in a different member of x's cohort.

For quantize, the cohort member is selected based on the preferred exponent for a result of that operation, whether or not the result is exact.

If the result's cohort does not include a member with the preferred exponent, the member with the exponent closest to the preferred exponent is used.

In the descriptions that follow, $Q(x)$ is the exponent q of the representation of a finite floating-point number x. If x is infinite, $Q(x)$ is $+\infty$.

5.3 Homogeneous general-computational operations

5.3.1 General operations

Implementations shall provide the following homogeneous general-computational operations for all supported floating-point formats available for arithmetic; all these operations never propagate non-canonical results. Their destination format is indicated as *sourceFormat*:

- *sourceFormat* **roundToIntegralTiesToEven**(*source*)
  sourceFormat **roundToIntegralTiesToAway**(*source*)
  sourceFormat **roundToIntegralTowardZero**(*source*)

sourceFormat **roundToIntegralTowardPositive**(*source*)
sourceFormat **roundToIntegralTowardNegative**(*source*)

See 5.9 for details.

The preferred exponent is max(Q(*x*), 0).

— *sourceFormat* **roundToIntegralExact**(*source*)

See 5.9 for details.

The preferred exponent is max(Q(*x*), 0).

— *sourceFormat* **nextUp**(*source*)
sourceFormat **nextDown**(*source*)

nextUp(*x*) is the least floating-point number in the format of *x* that compares greater than *x*. If *x* is the negative number of least magnitude in *x*'s format, **nextUp**(*x*) is −0. **nextUp**(±0) is the positive number of least magnitude in *x*'s format. **nextUp**(+∞) is +∞, and **nextUp**(−∞) is the finite negative number largest in magnitude. When *x* is NaN, then the result is according to 6.2. **nextUp**(x) is quiet except for *sNaNs*.

The preferred exponent is the least possible.

nextDown(*x*) is −nextUp(−*x*).

— *sourceFormat* **nextAfter**(*source, source*)

nextAfter(*x*, *y*) is the next floating-point number that neighbors *x* in the direction toward *y*, in the format of *x*:

— If either *x* or *y* is NaN, then the result is according to 6.2
— If $x=y$, then **nextAfter**(*x, y*) is canonicalized copySign(*x, y*)
— If $x<y$, then **nextAfter**(*x, y*) is nextUp(*x*); if $x>y$, then **nextAfter**(*x, y*) is nextDown(*x*)

Overflow is signaled when *x* is finite but **nextAfter**(*x*, *y*) is infinite; underflow is signaled when **nextAfter**(*x*, *y*) lies strictly between $\pm b^{emin}$; in both cases, inexact is signaled.

The preferred exponent is Q(*x*).

— *sourceFormat* **remainder**(*source, source*)

When $y\neq 0$, the remainder $r=$**remainder**(*x*, *y*) is defined for finite *x* and *y* regardless of the rounding-direction attribute by the mathematical relation $r=x-y\times n$, where *n* is the integer nearest the exact number x/y; whenever $|n-x/y| = ½$, then *n* is even. Thus, the remainder is always exact. If $r=0$, its sign shall be that of *x*. **remainder**(*x*, ∞) is *x* for finite *x*.

The preferred exponent is min(Q(*x*), Q(*y*)).

— *sourceFormat* **minNum**(*source, source*)
sourceFormat **maxNum**(*source, source*)
sourceFormat **minNumMag**(*source, source*)
sourceFormat **maxNumMag**(*source, source*)

minNum(*x*, *y*) is the canonicalized floating-point number *x* if $x<y$, *y* if $y<x$, the canonicalized floating-point number if one operand is a floating-point number and the other a quiet NaN. Otherwise it is either *x* or *y,* canonicalized. When either *x* or *y* is a signalingNaN, then the result is according to 6.2.

maxNum(*x*, *y*) is the canonicalized floating-point number *y* if $x<y$, *x* if $y<x$, the canonicalized floating-point number if one operand is a floating-point number and the other a quiet NaN. Otherwise it is either *x* or *y*, canonicalized. When either *x* or *y* is a signalingNaN, then the result is according to 6.2.

minNumMag(x, y) is the canonicalized floating-point number x if $|x|<|y|$, y if $|y|<|x|$, otherwise minNum(x, y).

maxNumMag(x, y) is the canonicalized floating-point number x if $|x|>|y|$, y if $|y|>|x|$, otherwise maxNum(x, y).

The preferred exponent is Q(x) if x is the result, Q(y) if y is the result.

5.3.2 Decimal operation

Implementations supporting decimal formats shall provide the following homogeneous general-computational operation for all supported decimal floating-point formats available for arithmetic; it never propagates non-canonical results. The destination format is indicated as *sourceFormat*:

— *sourceFormat* **quantize**(*source, source*)

For finite decimal operands x and y of the same format, **quantize**(x, y) is a floating-point number in the same format that has, if possible, the same numerical value as x and the same quantum as y. If the exponent is being increased, rounding according to the applicable rounding-direction attribute might occur: the result is a different floating-point representation and inexact is signaled if the result does not have the same numerical value as x. If the exponent is being decreased and the significand of the result would have more than p digits, invalid is signaled and the result is NaN. If one or both operands are NaN the rules in 6.2 are followed. Otherwise if only one operand is infinite then invalid is signaled and the result is NaN. If both operands are infinite then the result is canonical ∞ with the sign of x. **quantize** does not signal underflow or overflow.

The preferred exponent is Q(y).

5.3.3 *logBFormat* operations

Implementations shall provide the following general-computational operations for all supported floating-point formats available for arithmetic; these operations never propagate non-canonical floating-point results.

For each supported floating-point format available for arithmetic, languages define an associated *logBFormat* to contain the integral values of logB(x). The *logBFormat* shall have enough range to include all integers between $\pm 2 \times (emax + p)$ inclusive, which includes the scale factors for scaling between the finite numbers of largest and smallest magnitude.

If *logBFormat* is a floating-point format, then the following operations are homogeneous. If *logBFormat* is an integer format, then the first operand and the floating-point result of scaleB are of the same format.

— *sourceFormat* **scaleB**(*source, logBFormat*)

scaleB(x, N) is $x \times b^N$ for integral values N. The result is computed as if the exact product were formed and then rounded to the destination format, subject to the applicable rounding-direction attribute. When *logBFormat* is a floating-point format, the behavior of **scaleB** is language-defined when the second operand is non-integral. For non-zero values of N, **scaleB**(± 0, N) returns ± 0 and **scaleB**($\pm\infty$, N) returns $\pm\infty$. For zero values of N, **scaleB**(x, N) returns x.

The preferred exponent is Q(x) + N.

— *logBFormat* **logB**(*source*)

logB(x) is the exponent e of x, a signed integral value, determined as though x were represented with infinite range and minimum exponent. Thus $1 \le$ **scaleB**(x, $-$**logB**(x)) $< b$ when x is positive and finite. **logB**(1) is +0.

When *logBFormat* is a floating-point format, **logB**(NaN) is a NaN, **logB**(∞) is $+\infty$, and **logB**(0) is $-\infty$ and signals the divideByZero exception. When *logBFormat* is an integer format, then **logB**(NaN), **logB**(∞), and **logB**(0) return language-defined values outside the range $\pm 2 \times (emax + p - 1)$ and signal the invalid exception.

The preferred exponent is 0.

5.4 formatOf general-computational operations

5.4.1 Arithmetic operations

Implementations shall provide the following *formatOf* general-computational operations, for destinations of all supported floating-point formats available for arithmetic, and, for each destination format, for operands of all supported floating-point formats available for arithmetic with the same radix as the destination format. These operations never propagate non-canonical results.

— *formatOf*-**addition**(*source1, source2*)

The operation **addition**(x, y) computes $x+y$.

For inexact decimal results, the preferred exponent is the least possible. For exact decimal results, the preferred exponent is min(Q(x), Q(y)).

— *formatOf*-**subtraction**(*source1, source2*)

The operation **subtraction**(x, y) computes $x-y$.

For inexact decimal results, the preferred exponent is the least possible. For exact decimal results, the preferred exponent is min(Q(x), Q(y)).

— *formatOf*-**multiplication**(*source1, source2*)

The operation **multiplication**(x, y) computes $x\times y$.

For inexact decimal results, the preferred exponent is the least possible. For exact decimal results, the preferred exponent is Q(x)+Q(y).

— *formatOf*-**division**(*source1, source2*)

The operation **division**(x, y) computes x/y.

For inexact decimal results, the preferred exponent is the least possible. For exact decimal results, the preferred exponent is Q(x) − Q(y).

— *formatOf*-**squareRoot**(*source1*)

The operation **squareRoot**(x) computes $\sqrt{x}$. It has a positive sign for all operands ≥ 0, except that **squareRoot**(-0) shall be -0.

For inexact decimal results, the preferred exponent is the least possible. For exact decimal results, the preferred exponent is floor(Q(x)/2).

— *formatOf*-**fusedMultiplyAdd**(*source1, source2, source3*)

The operation **fusedMultiplyAdd**(x, y, z) computes $(x\times y)+z$ as if with unbounded range and precision, rounding only once to the destination format. No underflow, overflow, or inexact exception (see 7) can arise due to the multiplication, but only due to the addition; and so fusedMultiplyAdd differs from a multiplication operation followed by an addition operation.

For inexact decimal results, the preferred exponent is the least possible. For exact decimal results, the preferred exponent is min(Q(x)+Q(y), Q(z)).

— *formatOf*-**convertFromInt(***int*)

It shall be possible to convert from all supported signed and unsigned integer formats to all supported floating-point formats available for arithmetic. Integral values are converted exactly from integer formats to floating-point formats whenever the value is representable in both formats. If the converted value is not exactly representable in the destination format, the default result is determined according to the applicable rounding-direction attribute, and an inexact or floating-point overflow exception arises as specified in clause 7, just as with arithmetic operations. The signs of integer zeros are preserved. Integer zeros without signs are converted to +0.

The preferred exponent is 0.

Implementations shall provide the following *intFormatOf* general-computational operations for destinations of all supported integer formats and for operands of all supported floating-point formats available for arithmetic.

— *intFormatOf*-**convertToIntegerTiesToEven**(*source*)
intFormatOf-**convertToIntegerTowardZero**(*source*)
intFormatOf-**convertToIntegerTowardPositive**(*source*)
intFormatOf-**convertToIntegerTowardNegative**(*source*)
intFormatOf-**convertToIntegerTiesToAway**(*source*)

See 5.8 for details.

— *intFormatOf*-**convertToIntegerExactTiesToEven**(*source*)
intFormatOf-**convertToIntegerExactTowardZero**(*source*)
intFormatOf-**convertToIntegerExactTowardPositive**(*source*)
intFormatOf-**convertToIntegerExactTowardNegative**(*source*)
intFormatOf-**convertToIntegerExactTiesToAway**(*source*)

See 5.8 for details.

5.4.2 Conversion operations for all formats

Implementations shall provide the following *formatOf* conversion operations from all supported floating-point formats to all supported floating-point formats, including storage formats, as well as conversions to and from decimal character sequences. These operations never propagate non-canonical results. Some format conversion operations produce results in a different radix than the operands.

— *formatOf*-**convert**(*source*)

If the conversion is to a format in a different radix or to a narrower precision in the same radix, the result shall be rounded as specified in clause 4. Conversion to a format with the same radix but wider precision and range is always exact.

For inexact conversions from binary to decimal formats, the preferred exponent is the least possible. For exact conversions from binary to decimal formats, the preferred exponent is 0.

For conversions between decimal formats, the preferred exponent is Q(*source*).

— *formatOf*-**convertFromDecimalCharacter**(*decimalCharacterSequence*)

See 5.12 for details. The preferred exponent is Q(*decimalCharacterSequence*) which is the exponent value *q* of the last digit in the significand of the decimalCharacterSequence.

— *decimalCharacterSequence* **convertToDecimalCharacter**(*source, conversionSpecification*)

See 5.12 for details. The *conversionSpecification* specifies the precision and formatting of the *decimalCharacterSequence* result.

5.4.3 Conversion operations for binary formats

Implementations shall provide the following *formatOf* conversion operations to and from all supported binary floating-point formats, including storage formats; these operations never propagate non-canonical floating-point results.

— *formatOf*-**convertFromHexCharacter**(*hexCharacterSequence*)

See 5.12 for details.

— *hexCharacterSequence* **convertToHexCharacter**(*source, conversionSpecification*)

See 5.12 for details. The *conversionSpecification* specifies the precision and formatting of the *hexCharacterSequence* result.

5.5 Quiet-computational operations

5.5.1 Sign operations

Implementations shall provide the following homogeneous quiet-computational sign operations for all supported floating-point formats available for arithmetic; they only affect the sign. The operations treat floating-point numbers and NaNs alike, and signal no exception. They may propagate non-canonical encodings.

— *sourceFormat* **copy**(*source*)
sourceFormat **negate**(*source*)
sourceFormat **abs**(*source*)

copy(*x*) copies a floating-point operand *x* to a destination in the same format, with no change to the sign.

negate(*x*) copies a floating-point operand *x* to a destination in the same format, reversing the sign. $0 - x$ is not the same as $-x$ or **negate**(*x*).

abs(*x*) copies a floating-point operand *x* to a destination in the same format, changing the sign to positive.

— *sourceFormat* **copySign**(*source*, *source*)

copySign(*x, y)* copies a floating-point operand *x* to a destination in the same format as *x*, but with the sign of *y*.

5.5.2 Decimal re-encoding operations

For each supported decimal format (if any), the implementation shall provide the following operations to convert between the decimal format and the two standard encodings for that format. These operations enable portable programs that are independent of the implementation's encoding for decimal types to access data represented with either standard encoding. They may propagate non-canonical encodings.

— *decimalEncodingType* **encodeDecimal**(*decimalType*)
encodes the value of the operand using decimal encoding.
— *decimalType* **decodeDecimal**(*decimalEncodingType*)
decodes the decimal-encoded operand.
— *binaryEncodingType* **encodeBinary**(*decimalType*)
encodes the value of the operand using the binary encoding.
— *decimalType* **decodeBinary(***binaryEncodingType*)
decodes the binary-encoded operand.

where *decimalEncodingType* is a language-defined type for storing decimal-encoded decimal floating-point data, *binaryEncodingType* is a language-defined type for storing binary-encoded decimal floating-point data, and *decimalType* is the type of the given decimal floating-point format.

5.6 Signaling-computational operations

5.6.1 Comparisons

Implementations shall provide the following comparison operations, for all supported floating-point operands of the same radix in formats available for arithmetic:

— *boolean* **compareEqual**(*source1, source2*)
boolean **compareNotEqual**(*source1, source2*)
boolean **compareGreater**(*source1, source2*)
boolean **compareGreaterEqual**(*source1, source2*)
boolean **compareLess**(*source1, source2*)
boolean **compareLessEqual**(*source1, source2*)
boolean **compareSignalingNotGreater**(*source1, source2*)
boolean **compareSignalingLessUnordered**(*source1, source2*)
boolean **compareSignalingNotLess**(*source1, source2*)
boolean **compareSignalingGreaterUnordered**(*source1, source2*)
boolean **compareQuietGreater**(*source1, source2*)
boolean **compareQuietGreaterEqual**(*source1, source2*)
boolean **compareQuietLess**(*source1, source2*)
boolean **compareQuietLessEqual**(*source1, source2*)
boolean **compareUnordered**(*source1, source2*)
boolean **compareQuietNotGreater(***source1, source2*)
boolean **compareQuietLessUnordered**(*source1, source2*)
boolean **compareQuietNotLess(***source1, source2*)
boolean **compareQuietGreaterUnordered**(*source1, source2*)
boolean **compareOrdered(***source1, source2*).

See 5.11 for details.

5.6.2 Exception signaling

This operation signals the exceptions specified by its operand, invoking either default or, if explicitly requested, a language-defined alternate handling:

— *void* **signalException**(*exceptionGroupType*)
signals the exceptions specified in the *exceptionGroupType* operand, which can represent any subset of the exceptions.

The order in which the exceptions are signaled is unspecified..

5.7 Non-computational operations

5.7.1 Conformance predicates

Implementations shall provide the following non-computational operations, true if and only if the indicated conditions are true:

— *boolean* **is754version1985**(*void*)

is754version1985() is true if and only if this programming environment conforms to the earlier version of the standard.

— *boolean* **is754version2007**(*void*)

is754version2007() is true if and only if this programming environment conforms to this standard.

Implementations should make these predicates available at translation time (if applicable) in cases where their values can be determined at that point.

5.7.2 General operations

Implementations shall provide the following non-computational operations for all supported floating-point formats available for arithmetic. They are never exceptional, even for signaling NaNs.

— *enum* **class**(*source*)

class(*x*) tells which of the following ten classes *x* falls into:
signalingNaN
quietNaN
negativeInfinity
negativeNormal
negativeSubnormal
negativeZero
positiveZero
positiveSubnormal
positiveNormal
positiveInfinity.

— *boolean* **isSigned**(*source*)

isSigned(*x*) is true if and only if *x* has negative sign. **isSigned** applies to zeros and NaNs as well.

— *boolean* **isNormal**(*source*)

isNormal(*x*) is true if and only if *x* is normal (not zero, subnormal, infinite, or NaN).

— *boolean* **isFinite**(*source*)

isFinite(*x*) is true if and only if *x* is zero, subnormal or normal (not infinite or NaN).

— *boolean* **isZero**(*source*)

isZero(*x*) is true if and only if *x* is ±0.

— *boolean* **isSubnormal**(*source*)

isSubnormal(*x*) is true if and only if *x* is subnormal.

— *boolean* **isInfinite**(*source*)

isInfinite(*x*) is true if and only if *x* is infinite.

— *boolean* **isNaN**(*source*)

isNaN(*x*) is true if and only if *x* is a NaN.

— *boolean* **isSignaling**(*source*)

isSignaling(*x*) is true if and only if *x* is a signaling NaN.

— *boolean* **isCanonical**(*source*)

isCanonical(*x*) is true if and only if *x* is a finite number, infinity, or NaN that is canonical. Implementations should extend **isCanonical**(*x*) to formats which are not interchange formats in ways appropriate to those formats, which might, or might not, have finite numbers, infinities, or NaNs that are non-canonical.

— *int* **radix**(*source*)

radix(*x*) is the radix *b* of the format of *x*, that is, 2 or 10.

— *boolean* **totalOrder**(*source, source*)

totalOrder(*x, y*) is defined in 5.10.

— *boolean* **totalOrderMag**(*source, source*)

totalOrderMag(*x, y*) is totalOrder(abs(*x*), abs(*y*)).

5.7.3 Decimal operation

Implementations supporting decimal formats shall provide the following non-computational operation for all supported decimal floating-point formats available for arithmetic:

— *boolean* **sameQuantum**(*source, source*)

For numerical decimal operands *x* and *y* of the same format, **sameQuantum**(*x*, *y*) is true if the exponents of *x* and *y* are the same, that is, Q(*x*) = Q(*y*), and false otherwise. **sameQuantum**(NaN, NaN) and **sameQuantum**(∞, ∞) are true; if exactly one operand is infinite or exactly one operand is NaN, **sameQuantum** is false. **sameQuantum** signals no exception.

5.7.4 Operations on subsets of flags

Implementations shall provide the following non-computational operations that act upon multiple status flags collectively:

— *void* **lowerFlags**(*exceptionGroupType*)

lowers (clears) the flags corresponding to the exceptions specified in the *exceptionGroupType* operand, which can represent any subset of the exceptions.

— *boolean* **testFlags**(*exceptionGroupType*)

queries whether any of the flags corresponding to the exceptions specified in the *exception-GroupType* operand, which can represent any subset of the exceptions, are raised.

— *boolean* **testSavedFlags**(*flagsType, exceptionGroupType*)

queries whether any of the flags in the *flagsType* operand corresponding to the exceptions specified in the *exceptionGroupType* operand, which can represent any subset of the exceptions, are raised.

— *void* **restoreFlags**(*flagsType*, *exceptionGroupType*)

restores the flags corresponding to the exceptions specified in the *exceptionGroupType* operand, which can represent any subset of the exceptions, to their state represented in the *flagsType* operand.

— *flagsType* **saveFlags**(*exceptionGroupType*)

returns a representation of the state of those flags corresponding to the exceptions specified in the *exceptionGroupType* operand.

The return value of the saveFlags operation is for use as the first operand to a restoreFlags or testSavedFlags operation in the same program; this standard does not require support for any other use.

5.8 Details of conversions from floating-point to integer formats

Implementations shall provide conversion operations from all supported floating-point formats available for arithmetic to all supported signed and unsigned integer formats. Integral values are converted exactly from floating-point formats to integer formats whenever the value is representable in both formats.

Conversion to integer shall be effected by rounding as specified in clause 4, but the rounding direction is indicated by the operation name.

When a NaN or infinite operand cannot be represented in the destination format and this cannot otherwise be indicated, the invalid exception shall be signaled. When a numeric operand would convert to an integer outside the range of the destination format, the invalid exception shall be signaled if this situation cannot otherwise be indicated.

When the rounded-to-integral value of the conversion operation's result differs from its operand value, yet is representable in the destination format, the inexact exception might be signaled in certain circumstances and not others.

The inexact exception should be signaled if an inexact conversion was invoked by a language's rules for implicit conversions or expressions involving mixed types.

The operations for conversion from floating-point to a specific signed or unsigned integer format without signaling inexact are:

— *intFormatOf*-**convertToIntegerTiesToEven**(*x*)
rounds *x* to the nearest integral value, with halfway cases rounded to even
— *intFormatOf*-**convertToIntegerTowardZero**(*x*)
rounds *x* to an integral value toward zero
— *intFormatOf*-**convertToIntegerTowardPositive**(*x*)
rounds *x* to an integral value toward positive infinity
— *intFormatOf*-**convertToIntegerTowardNegative**(*x*)
rounds *x* to an integral value toward negative infinity
— *intFormatOf*-**convertToIntegerTiesToAway**(*x*)
rounds *x* to the nearest integral value, with halfway cases rounded away from zero.

The operations for conversion from floating-point to a specific signed or unsigned integer format, signaling if inexact, are:

— *intFormatOf*-**convertToIntegerExactTiesToEven**(*x*)
rounds *x* to the nearest integral value, with halfway cases rounded to even
— *intFormatOf*-**convertToIntegerExactTowardZero**(*x*)
rounds *x* to an integral value toward zero
— *intFormatOf*-**convertToIntegerExactTowardPositive**(*x*)
rounds *x* to an integral value toward positive infinity
— *intFormatOf*-**convertToIntegerExactTowardNegative**(*x*)
rounds *x* to an integral value toward negative infinity
— *intFormatOf*-**convertToIntegerExactTiesToAway**(*x*)
rounds *x* to the nearest integral value, with halfway cases rounded away from zero.

5.9 Details of operations to round a floating-point datum to integral value

Several operations round a floating-point number to an integral valued floating-point number in the same format.

The rounding is analogous to that specified in clause 4, but the rounding chooses only from among those floating-point numbers of integral values in the format. These operations convert zero operands to zero results of the same sign, and infinite operands to infinite results of the same sign.

For the following operations, the rounding direction is implied by the operation name and does not depend on a rounding-direction attribute. These operations do not signal any exception except for signaling NaN input.

— *sourceFormat* **roundToIntegralTiesToEven**(*x*)
rounds *x* to the nearest integral value, with halfway cases rounding to even

— *sourceFormat* **roundToIntegralTowardZero**(*x*)
rounds *x* to an integral value toward zero

— *sourceFormat* **roundToIntegralTowardPositive**(*x*)
rounds *x* to an integral value toward positive infinity

— *sourceFormat* **roundToIntegralTowardNegative**(*x*)
rounds *x* to an integral value toward negative infinity

— *sourceFormat* **roundToIntegralTiesToAway**(*x*)
rounds *x* to the nearest integral value, with halfway cases rounding away from zero.

For the following operation, the rounding direction is the applicable rounding-direction attribute. This operation signals invalid for signaling NaN, and for a numerical operand, signals inexact if the result does not have the same numerical value as *x*.

— *sourceFormat* **roundToIntegralExact**(*x*)
rounds *x* to an integral value according to the applicable rounding-direction attribute.

5.10 Details of totalOrder predicate

For each supported floating-point format available for arithmetic, an implementation shall provide the following predicate which defines an ordering among all operands in a particular format.

totalOrder(*x, y*) imposes a total ordering on canonical members of the format of *x* and *y*:

a) if $x < y$, totalOrder(*x, y*) is true
b) if $x > y$, totalOrder(*x, y*) is false
c) if $x = y$:
    1) totalOrder(−0, +0) is true
    2) totalOrder(+0, −0) is false
    3) if *x* and *y* represent the same floating-point datum:
        i) if *x* and *y* have negative sign,
        totalOrder(*x*, *y*) is true if and only if the exponent of $x \geq$ the exponent of *y*
        ii) otherwise
        totalOrder(*x*, *y*) is true if and only if the exponent of $x \leq$ the exponent of *y*.

    Note that totalOrder does not impose a total ordering on all encodings in a format. In particular, it does not distinguish among different encodings of the same floating-point representation, as when one or both encodings are non-canonical.
d) if *x* and *y* are unordered numerically because *x* or *y* is NaN:
    1) totalOrder(−NaN, *y*) is true where −NaN represents a NaN with negative sign bit and *y* is a floating-point number
    2) totalOrder(*x*, +NaN) is true where +NaN represents a NaN with positive sign bit and *x* is a floating-point number
    3) if *x* and *y* are both NaNs, then totalOrder reflects a total ordering based on
        i) negative sign is lower than positive sign
        ii) signaling is lower than quiet for +NaN, reverse for −NaN
        iii) lesser payload is lower than greater payload for +NaN, reverse for −NaN.

Neither signaling nor quiet NaNs signal an exception.

For canonical *x* and *y*, totalOrder(*x, y*) and totalOrder(*y, x*) are both true only if *x* and *y* are bitwise identical.

5.11 Details of comparison predicates

For every supported floating-point format available for arithmetic, it shall be possible to compare one floating-point datum to another in that format. Additionally, floating-point data represented in different formats shall be comparable as long as the operands' formats have the same radix.

Comparisons are exact and never overflow or underflow. Four mutually exclusive relations are possible: *less than*, *equal*, *greater than*, and *unordered*. The last case arises when at least one operand is NaN. Every NaN shall compare *unordered* with everything, including itself. Comparisons shall ignore the sign of zero (so +0 = −0). Infinite operands of the same sign shall compare *equal*.

Languages define how the result of a comparison shall be delivered, in one of two ways: either as a relation identifying one of the four relations listed above, or as a true-false response to a predicate that names the specific comparison desired.

Table 10, Table 11, and Table 12 exhibit twenty functionally distinct useful predicates and negations with various ad-hoc and traditional names and symbols. Each predicate is true if any of its indicated relations is true. The relation "?" indicates an *unordered* relation. Table 11 lists four unordered-signaling predicates and their negations that cause an invalid operation exception when the relation is unordered. That invalid exception defends against unexpected quiet NaNs arising in programs written using the standard predicates {<, <=, >=, >} and their negations, without considering the possibility of a quiet NaN operand. Programs that explicitly take account of the possibility of quiet NaN operands may use the unordered-quiet predicates in Table 12 which do not signal such an invalid exception.

Note that predicates come in pairs, each a logical negation of the other; applying a prefix such as NOT to negate a predicate in Table 10, Table 11, and Table 12 reverses the true/false sense of its associated entries, but does not change whether *unordered* relations cause an invalid operation exception.

The unordered-quiet predicates in Table 10 do not signal an exception on quiet NaN operands:

Table 10—Required unordered-quiet predicate and negation

Unordered-quiet predicate		Unordered-quiet negation	
True relations	**Names**	**True relations**	**Names**
EQ	compareEqual =	LT GT UN	compareNotEqual ?<> NOT(=) ≠

The unordered-signaling predicates in Table 11, intended for use by programs *not* written to take into account the possibility of NaN operands, signal an invalid exception on quiet NaN operands:

Table 11—Required unordered-signaling predicates and negations

Unordered-signaling predicate		Unordered-signaling negation	
True relations	**Names**	**True relations**	**Names**
GT	compareGreater >	EQ LT UN	compareSignalingNotGreater NOT(>)
GT EQ	compareGreaterEqual >= ≥	LT UN	compareSignalingLessUnordered NOT(>=)
LT	compareLess <	EQ GT UN	compareSignalingNotLless NOT(<)
LT EQ	compareLessEqual <= ≤	GT UN	compareSignalingGreaterUnordered NOT(<=)

The unordered-quiet predicates in Table 12, intended for use by programs written to take into account the possibility of NaN operands, do not signal an exception on quiet NaN operands:

Table 12—Required unordered-quiet predicates and negations

Unordered-quiet predicate		Unordered-quiet negation	
True relations	**Names**	**True relations**	**Names**
GT	CompareQuietGreater isGreater	EQ LT UN	compareQuietNotGreater ?<=
GT EQ	compareQuietGreaterEqual isGreaterEqual	LT UN	compareQuietLessUnordered ?<
LT	compareQuietLess isLess	EQ GT UN	compareQuietNotLess ?>=
LT EQ	compareQuietLessEqual isLessEqual	GT UN	compareQuietGreaterUnordered ?>
UN	compareUnordered ? isUnordered	LT EQ GT	compareOrdered <=> NOT(?)

There are two ways to write the logical negation of a predicate, one using NOT explicitly and the other reversing the relational operator. Thus in programs written without considering the possibility of a NaN operand, the logical negation of the unordered-signaling predicate (X < Y) is just the unordered-signaling predicate NOT(X < Y); the unordered-quiet reversed predicate (X ?>= Y) is different in that it does not signal an invalid operation exception when X and Y are *unordered*. In contrast, the logical negation of (X = Y) might be written either NOT(X = Y) or (X ?<> Y); in this case both expressions are functionally equivalent to (X ≠ Y).

5.12 Details of conversion between floating-point data and external character sequences

This clause specifies conversions between supported formats and external character sequence formats. Note that conversions between supported formats of different radices are correctly rounded and set exceptions correctly as described in 5.4.2, subject to limits stated in clause 5.12.2 below.

Implementations shall provide conversions between each supported binary format and external decimal character sequences such that, under roundTiesToEven, conversion from the supported format to external decimal character sequence and back recovers the original floating-point representation, except that a signaling NaN might be converted to a quiet NaN. See 5.12.1and 5.12.2 for details.

Implementations shall provide exact conversions from each supported decimal format to external decimal character sequences, and shall provide conversions back that recover the original floating-point representation, except that a signaling NaN might be converted to a quiet NaN. See 5.12.1and 5.12.2 for details.

Implementations shall provide exact conversions from each supported binary format to external character sequences representing numbers with hexadecimal digits for the significand, and shall provide conversions back that recover the original floating-point representation, except that a signaling NaN might be converted to a quiet NaN. See 5.12.1 and 5.12.3 for details.

This clause primarily discusses conversions during program execution; there is one special consideration applicable to program translation separate from program execution: translation-time conversion of constants in program text from external character sequences to supported formats, in the absence of other specification in the program text, shall use this standard's default rounding direction and language-defined exception handling. An implementation might also provide means to permit constants to be translated at execution time with the attributes in effect at execution time and exceptions generated at execution time.

Issues of character codes (ASCII, Unicode, etc.) are not defined by this standard.

5.12.1 External character sequences representing zeros, infinities, and NaNs

The conversions (described in 5.4.2) from supported formats to external character sequences and back that recover the original floating-point representation, shall recover zeros, infinities, and quiet NaNs, as well as non-zero finite numbers. In particular, signs of zeros and infinities are preserved.

Conversion of an infinity in a supported format to an external character sequence shall produce a language-defined one of "inf" or "infinity" or a sequence that is equivalent except for case (*e.g.*, "Infinity" or "INF"), with a preceding minus sign if the input is negative. Whether the conversion produces a preceding plus sign if the input is positive is language-defined.

Conversion of external character sequences "inf" and "infinity", regardless of case, with an optional preceding sign, to a supported floating-point format shall produce an infinity (with the same sign as the input).

Conversion of a quiet NaN in a supported format to an external character sequence shall produce a language-defined one of "nan" or a sequence that is equivalent except for case (*e.g.*, "NaN"), with an optional preceding sign. The sign of a NaN has no meaning.

Conversion of a signaling NaN in a supported format to an external character sequence should produce a language-defined one of "snan" or "nan" or a sequence that is equivalent except for case, with an optional preceding sign. If the conversion of a signaling NaN produces "nan" or a sequence that is equivalent except for case, with an optional preceding sign, then the invalid exception should be signaled.

Conversion of external character sequences "nan", regardless of case, with an optional preceding sign, to a supported floating-point format shall produce a quiet NaN.

Conversion of an external character sequence "snan", regardless of case, with an optional preceding sign, to a supported format should either produce a signaling NaN or else produce a quiet NaN and signal the invalid exception.

Language standards should provide an optional conversion of NaNs in a supported format to external character sequences which appends to the basic NaN character sequences a suffix that can represent the NaN payload (see 6.2). The form and interpretation of the payload suffix is language-defined. The language standard shall require that any such optional output sequences be accepted as input in conversion of external character sequences to supported formats.

5.12.2 External decimal character sequences representing finite numbers

An implementation shall provide operations which convert from all supported floating-point formats to external decimal character sequences (see 5.4.2). For finite numbers, these operations might be thought of as parameterized by the source format, the number of significant digits in the result (if specified), and whether the quantum is preserved (for decimal formats). Note that specifying the number of significant digits and specifying quantum preservation are mutually incompatible. The means of specifying the number of significant digits and of specifying quantum preservation are language-defined and are typically embodied in the *conversionSpecification* of clause 5.4.2.

An implementation shall also provide operations which convert external decimal character sequences to all supported formats. These operations might be thought of as parameterized by the result format.

Within the limits stated in this clause, conversions in both directions shall preserve the value of a number unless rounding is necessary and shall preserve its sign. If rounding is necessary, they shall use correct rounding and shall correctly signal inexact and other exceptions.

All conversions from external character sequences to supported decimal formats shall preserve the quantum unless rounding is necessary. At least one conversion from each supported decimal format shall preserve the quantum as well as the value and sign.

If a conversion to an external character sequence requires an exponent but the exponent is not of sufficient width to avoid overflow or underflow (see 7.4 and 7.5), the overflow or underflow should be indicated to the user by appropriate language-defined character sequences.

For the purposes of discussing the limits on correctly rounded conversion, define the following quantities:

- for binary16, *Pmin*(binary16) = 5
- for binary32, *Pmin*(binary32) = 9
- for binary64, *Pmin*(binary64) = 17
- for binary128, *Pmin*(binary128) = 36
- for all other binary formats *bf*, $Pmin(bf) = 1 + \text{ceiling}(p \times \log_{10}(2))$, where p is the number of significant bits in *bf*
- M = max(*Pmin*(*bf*)) for all supported binary formats *bf*
- for decimal32, *Pmin*(decimal32) = 7
- for decimal64, *Pmin*(decimal64) = 16
- for decimal128, *Pmin*(decimal128) = 34
- for all other decimal formats *df*, *Pmin*(*df*) is the number of significant digits in *df*.

Conversions to and from supported decimal formats shall be correctly rounded regardless of how many digits are requested or given.

There might be an implementation-defined limit on the number of significant digits that can be converted with correct rounding to and from supported binary formats. That limit, H, shall be such that $H \geq M + 3$ and it should be that H is unbounded.

For all supported binary formats the conversion operations shall support correctly rounded conversions to or from external character sequences for all significant digit counts from 1 through H (that is, for all expressible counts if H is unbounded).

Conversions from supported binary formats to external character sequences for which more than H significant digits are specified shall pad with trailing zeros.

Conversion from a character sequence of more than H significant digits or larger in exponent range than the destination binary format first shall be correctly rounded to H digits according to the applicable rounding direction and shall signal exceptions as though narrowing from a wider format and then the resulting character sequence of H digits shall be converted with correct rounding according to the applicable rounding direction.

As a consequence of the foregoing, the following are true:

- conversions to or from decimal formats are correctly rounded
- for binary formats, all conversions of H significant digits or fewer round correctly according to the applicable rounding direction; conversions of greater than H significant digits might incur ~~an~~ additional rounding of the order of $10^{(M-H)} < 10^{-3}$ units in the last place
- intervals are respected, in the sense that directed-rounding constraints are honored even when more than H significant digits are given: directed rounding has the correct sign in all cases, and the error never exceeds 1.001 units in the last place in magnitude
- conversions are monotonic; increasing the value of a supported floating-point number does not decrease its value after conversion to an external character sequence, and increasing the value of an external character sequence does not decrease its value after conversion to a supported floating-point number
- conversions from a supported binary format *bf* to an external character sequence and back again results in a copy of the original number so long as there are at least *Pmin*(*bf*) significant digits specified
- conversions from a supported decimal format *df* to an external character sequence and back again results in a canonical copy of the original number so long as the conversion to the external character sequence is one that preserves the quantum
- conversions from a supported decimal format *df* to an external character sequence and back again recovers the value (but not necessarily the quantum) of the original number so long as there are at least *Pmin*(*df*) significant digits specified
- all implementations exchange equivalent decimal sequences: two decimal character sequences are equivalent if they represent the same value (and quantum, for decimal formats); if two implementations support a given format they convert any floating-point representation in that format to equivalent decimal character sequences when the same number of digits is specified and (for binary formats) the specified number of digits is no greater than H, or (for decimal formats) when the quantum-preserving conversion is specified
- similarly, any two implementations convert equivalent decimal sequences to the same floating-point number (with the same quantum, for decimal formats) when the number of significant digits and the result format are supported on both implementations.

NOTE—H should be as large as practical, noting that "practical" might well include "unbounded" on many systems because any H at least as large as the number of digits required for the longest exact decimal representation is effectively as good as unbounded. The length of the longest exact decimal representation is less than 12,000 digits for binary128.

5.12.3 External hexadecimal character sequences representing finite numbers

Implementations supporting binary formats shall provide conversions between all supported binary formats and external hexadecimal character sequences.

External hexadecimal character sequences for finite numbers are described by the following grammar which defines a *hexSequence*:

```
sign            [+−]
digit           [0123456789]
hexDigit        [0123456789abcdefABCDEF]
hexExpIndicator [Pp]
hexIndicator    "0" [Xx]
hexSignificand  ( {hexDigit}* "." {hexDigit}+ | {hexDigit}+ "." | {hexDigit}+ )
hexExponent     {hexExpIndicator} {sign}? {digit}+
hexSequence     {sign}? {hexIndicator} {hexSignificand} {hexExponent}
```

where each line is a name followed by a rule in which '[...]' selects one of the terminal characters listed between the brackets, '{...}' refers to an earlier named rule, '(...|...|...)' indicates a choice of one of three alternatives, straight double quotes enclose a terminal character, '?' indicates that there shall be either no instance or one instance of the preceding item, '*' indicates that there shall be zero or more instances of the preceding item, and '+' indicates that there shall be one or more instances of the preceding item.

The *hexSignificand* is interpreted as a hexadecimal constant in which each *hexDigit* represents a value in the range 0 through 15 with the letters 'a' through 'f' representing 10 through 15, regardless of case. Within the *hexSignificand*, the first (leftmost) character is the most significant. If present, the period defines the start of a hexadecimal fractional part; if the period is to the right of all hexadecimal digits the *hexSignificand* is an integer. The *hexExponent* is interpreted as an optionally-signed integer expressed in decimal following the *hexExpIndicator*, again with the most significant digit first.

The value of a *hexSequence* is the value of the *hexSignificand* multiplied by two raised to the power of the value of the *hexExponent*, negated if there is a leading '−' sign. The *hexIndicator* and the *hexExpIndicator* have no effect on the value.

When converting to hexadecimal character sequences in the absence of an explicit precision specification, enough hexadecimal characters shall be used to represent the binary floating-point number exactly. Conversions to hexadecimal character sequences with an explicit precision specification, and conversions from hexadecimal character sequences to supported binary formats, are correctly rounded according to the applicable binary rounding-direction attribute, and signals all exceptions appropriately.

NOTE — The external hexadecimal character sequences described here follow those specified for finite numbers in ISO/IEC 9899, Second Edition 1999–12–01, Programming languages — C (C99), in clauses:

- 6.4.4.2 floating constants
- 7.19.6.1 fprintf (conversion specifiers 'a' and 'A')
- 7.19.6.2 fscanf (conversion specifier 'a')
- 7.20.1.3 strtod.

6. Infinity, NaNs, and sign bit

6.1 Infinity arithmetic

The behavior of infinity in floating-point arithmetic is derived from the limiting cases of real arithmetic with operands of arbitrarily large magnitude, when such a limit exists. Infinities shall be interpreted in the affine sense, that is:

$$-\infty < \{\text{every finite number}\} < +\infty.$$

Operations on infinite operands are usually exact and therefore signal no exceptions. The exceptions that do pertain to infinities are signaled only when:

- ∞ is an invalid operand (see 7.2)
- ∞ is created from finite operands by overflow (see 7.4) or division by zero (see 7.3)
- remainder(subnormal, ∞) signals underflow
- nextAfter(x, ∞) signals underflow and inexact if the result would be subnormal
- nextAfter(maximum normal, ∞) signals overflow and inexact if the result would be infinite

and similarly for negative values.

6.2 Operations with NaNs

Two different kinds of NaN, signaling and quiet, shall be supported in all floating-point operations. Signaling NaNs afford representations for uninitialized variables and arithmetic-like enhancements (such as complex-affine infinities or extremely wide range) that are not the subject of the standard. Quiet NaNs should, by means left to the implementer's discretion, afford retrospective diagnostic information inherited from invalid or unavailable data and results. To facilitate propagation of diagnostic information contained in NaNs, as much of that information as possible should be preserved in NaN results of computational operations.

Under default exception handling, any operation signaling an invalid exception and for which a floating-point result is to be delivered shall deliver a quiet NaN.

Signaling NaNs shall be reserved operands that, under default exception handling, signal the invalid operation exception (see 7.1) for every general-computational and signaling-computational operation except for the conversions described in 5.12. For non-default treatment see 8)

Every general-computational and quiet-computational operation involving one or more input NaNs, none of them signaling, shall signal no exception, except fusedMultiplyAdd might signal the invalid operation exception (see 7.2). For an operation with quiet NaN inputs, other than maximum and minimum operations, if a floating-point result is to be delivered the result shall be a quiet NaN which should be one of the input NaNs. If the trailing significand field of a decimal input NaN is canonical then the bit pattern of that field shall be preserved if that NaN is chosen as the result NaN. Note that format conversions, including conversions between supported formats and external representations as character sequences, might be unable to deliver the same NaN. Quiet NaNs signal exceptions on some operations that do not deliver a floating-point result; these operations, namely comparison and conversion to a format that has no NaNs, are discussed in 5.6, 5.8, and 7.2.

6.2.1 NaN encodings in binary formats

This subclause further specifies the encodings of NaNs as bit strings when they are the results of operations. When encoded, all NaNs have a sign bit and a pattern of bits necessary to identify the encoding as a NaN and which determines its kind (sNaN vs. qNaN). The remaining bits, which are in the trailing field, encode the payload, which might be diagnostic information (see above).

All binary NaN bit strings have all the bits of the biased exponent field E set to 1 (see 3.4). A quiet NaN bit string should be encoded with the first bit (d_1) of the trailing significand field T being 1. A signaling NaN bit string should be encoded with the first bit of the trailing significand field being 0. If the first bit of the

trailing significand is 0, some other bit of the trailing significand field must be non-zero to distinguish the NaN from infinity. In the preferred encoding just described, a signaling NaN shall be quieted by setting d_1 to 1, leaving the remaining bits of T unchanged.

For binary formats, the payload is the $p-2$ least significant bits of the trailing significand field.

6.2.2 NaN encodings in decimal formats

A decimal signaling NaN shall be quieted by clearing G_5 and leaving the values of the digits d_1 through d_{p-1} of the trailing significand unchanged (see 3.5).

Any computational operation which produces, propagates, or quiets a decimal format NaN shall set the bits G_6 through G_{w+4} of G to 0, and shall generate only a canonical trailing significand field.

For decimal formats, the payload is the trailing significand field.

6.2.3 NaN propagation

An operation which propagates a NaN operand to its result and has a single NaN as an input should produce a NaN with the payload of the input NaN.

If two or more inputs are NaN, then the payload of the resulting NaN should be identical to the payload of one of the input NaNs. This standard does not specify which of the input NaNs will provide the payload.

Conversion of a quiet NaN from a narrower format to a wider format in the same radix, and then back to the same narrower format, should not change the quiet NaN payload in any way except to make it canonical.

Conversion of a quiet NaN to a floating-point format of the same or a different radix that does not allow the payload to be preserved, shall return a quiet NaN that should provide some language-defined diagnostic information.

There should be means to read and write NaN payloads from and to external character sequences (see 5.12.1).

6.3 The sign bit

When either an input or result is NaN, this standard does not interpret the sign of a NaN. Note, however, that operations on bit strings—copy, negate, abs, copySign—specify the sign bit of a NaN result, sometimes based upon the sign bit of a NaN operand. The logical predicate totalOrder is also affected by the sign bit of a NaN operand. For all other operations, this standard does not specify the sign bit of a NaN result, even when there is only one input NaN, or when the NaN is produced from an invalid operation.

When neither the inputs nor result are NaN, the sign of a product or quotient is the exclusive OR of the operands' signs; the sign of a sum, or of a difference $x-y$ regarded as a sum $x+(-y)$, differs from at most one of the addends' signs; and the sign of the result of conversions, the quantize operation, the roundToIntegral operations, and the roundToIntegralExact (see 5.3.1) is the sign of the operand. These rules shall apply even when operands or results are zero or infinite.

When the sum of two operands with opposite signs (or the difference of two operands with like signs) is exactly zero, the sign of that sum (or difference) shall be +0 in all rounding-direction attributes except roundTowardNegative; under that attribute, the sign of an exact zero sum (or difference) shall be −0. However, $x+x = x-(-x)$ retains the same sign as x even when x is zero.

When $(a \times b)+c$ is exactly zero, the sign of fusedMultiplyAdd(a, b, c) shall be determined by the rules above for a sum of operands. When the exact result of $(a \times b)+c$ is non-zero yet the result of fusedMultiplyAdd is zero because of rounding, the zero result takes the sign of the exact result.

Except that squareRoot(−0) shall be −0, every valid squareRoot result shall have a positive sign.

7. Default exception handling

7.1 Overview: exceptions and flags

There are five types of exceptions that shall be signaled. This clause specifies default nonstop exception handling, which entails raising a status flag (except as stated below), delivering a default result, and continuing execution. A language standard might define attributes for alternate exception handling and means for programmers to invoke them.

For each type of exception the implementation shall provide a status flag that shall be raised when the corresponding exception is signaled. It shall be lowered only at the user's request. The user shall be able to test and to alter the status flags individually or collectively, and shall further be able to save and restore all five at one time (see 5.7.4).

A program that does not inherit status flags from another source, begins execution with all status flags lowered.

Language standards should specify defaults in the absence of any explicit program specification, governing:

- whether any particular flag exists (in the sense of being testable by non-programmatic means such as debuggers) outside of scopes in which a program explicitly sets or tests that flag
- when flags have scope greater than within an invoked function, whether and when an asynchronous event, such as raising or lowering it in another thread or signal handler, affects the flag tested within that invoked function
- when flags have scope greater than within an invoked function, whether a flag's state can be determined by non-programmatic means (such as a debugger) within that invoked function
- whether flags raised in invoked functions set flags in invoking functions
- whether flags raised in invoking functions set flags in invoked functions
- whether to allow, and if so the means, to specify that flags shall be persistent in the absence of any explicit program statement otherwise:
  - the flags standing at the beginning of execution of a particular function are inherited from an outer environment, typically an invoking function
  - on return from or termination of an invoked function, the flags standing in an invoking function are the flags that were standing in the function at the time of return or termination.

An invocation of the signalException operation of 5.6.2 might signal any combination of exceptions. For an invocation of any other operation required by this standard, at most two exceptions might be signaled, in just these combinations: overflow followed by inexact, and underflow followed by inexact.

The inexact exception is signaled if the overflow exception receives default handling, and might be signaled if the underflow exception receives default handling (see 7.5).

In general, when an operation signals more than one exception, none of which have alternate exception handling enabled, each signaled exception will receive its default handling.

When an operation signals more than one exception, some or all of which have alternate exception handling enabled, alternate exception handling will be invoked for the most important exception, and languages define whether other signaled exceptions receive default handling, alternate handling, or are ignored. Exceptions are listed in this clause in order of decreasing importance (invalid, divideByZero, overflow, underflow, inexact).

For the computational operations defined in this standard, exceptions are defined below to be signaled if and only if certain conditions arise. That is not meant to imply whether those exceptions are signaled by operations not specified by this standard such as complex arithmetic or certain transcendental functions. Those and other operations, not specified by this standard, should signal those exceptions according to the definitions below for standard operations, but that might not always be economical. Standard exceptions for nonstandard functions are language-defined.

7.2 Invalid operation

The invalid operation exception is signaled if and only if there is no usefully definable result. In these cases the operands are invalid for the operation to be performed.

For operations producing results in floating-point format, the default result of an invalid exception operation shall be a quiet NaN that should provide some diagnostic information (see 6.2). These invalid exception operations are:

a) any general-computational or signaling-computational operation on a signaling NaN (see 6.2), except for some conversions (see 5.12)
b) multiplication: multiplication(0, ∞) or multiplication(∞, 0)
c) fusedMultiplyAdd: fusedMultiplyAdd(0, ∞, *c*) or fusedMultiplyAdd(∞, 0, *c*) unless *c* is a quiet NaN; if *c* is a quiet NaN then it is implementation defined whether the invalid operation exception is signaled
d) addition or subtraction or fusedMultiplyAdd: magnitude subtraction of infinities, such as: addition(+∞, −∞)
e) division: division(0, 0) or division(∞, ∞)
f) remainder: remainder(*x, y*), when *y* is zero or *x* is infinite and neither is NaN
g) squareRoot if the operand is less than zero
h) quantize when the result does not fit in the destination format or when one operand is finite and the other is infinite

For operations producing no result in floating-point format, the invalid exception operations are:

i) conversion of a floating-point number to an integer format, when the source is NaN, infinity, or a value which would convert to an integer outside the range of the result format under the applicable rounding attribute
j) comparison by way of unordered-signaling predicates listed in Table 11, when the operands are *unordered*
k) logB(NaN), logB(∞), or logB(0) when *logBFormat* is an integer format (see 5.3.3).

7.3 Division by zero

The divideByZero exception shall be signaled if and only if an exact infinite result is defined for an operation on finite operands. The divideByZero exception shall be signaled if the divisor is zero and the dividend is a finite non-zero number. The default result shall be a correctly signed ∞ (see 6.3).

Also, when *logBFormat* is a floating-point format, logB(0) signals the divideByZero exception with default result −∞.

7.4 Overflow

The overflow exception shall be signaled if and only if the destination format's largest finite number is exceeded in magnitude by what would have been the rounded floating-point result (see clause 4) were the exponent range unbounded. The default result shall be determined by the rounding-direction attribute and the sign of the intermediate result as follows:

a) roundTiesToEven and roundTiesToAway carries all overflows to ∞ with the sign of the intermediate result
b) roundTowardZero carries all overflows to the format's largest finite number with the sign of the intermediate result
c) roundTowardNegative carries positive overflows to the format's largest finite number, and carries negative overflows to −∞
d) roundTowardPositive carries negative overflows to the format's most negative finite number, and carries positive overflows to +∞.

In addition, under default exception handling for overflow, the overflow and inexact exceptions shall be signaled.

However nextAfter(x, y) signals overflow and inexact if and only if nextAfter is infinite and differs from the finite number x.

7.5 Underflow

The underflow exception shall be signaled when a tiny non-zero result is detected. For binary formats, this shall be either:

a) *After rounding*—when a non-zero result computed as though the exponent range were unbounded would lie strictly between $\pm b^{emin}$, or

b) *Before rounding*—when a non-zero result computed as though both the exponent range and the precision were unbounded would lie strictly between $\pm b^{emin}$.

The implementer shall choose how tininess is detected, but shall detect tininess in the same way for all operations in radix two, including conversion operations under a binary rounding attribute.

For decimal formats, tininess is detected before rounding—when a non-zero result computed as though both the exponent range and the precision were unbounded would lie strictly between $\pm b^{emin}$.

Thee default exception handling for underflow shall always deliver a rounded result. The method for detecting tininess does not affect the rounded result delivered, which might be zero, subnormal, or $\pm b^{emin}$.

In addition, under default exception handling for underflow, if the rounded result is inexact—that is, it differs from what would have been computed were both exponent range and precision unbounded—the underflow flag shall be raised and the inexact (see 7.6) exception shall be signaled. If the rounded result is exact, no flag is raised and no Inexact exception is signaled.

However, nextAfter(x, y) signals underflow and inexact if and only if the result is strictly between $\pm b^{emin}$ and compares unequal to x.

7.6 Inexact

If the rounded result of an operation is inexact—that is, it differs from what would have been computed were both exponent range and precision unbounded—then the inexact exception shall be signaled. The rounded or overflowed result shall be delivered to the destination.

nextAfter(x, y) signals inexact if and only if nextAfter also signals overflow or underflow.

8. Alternate exception handling attributes

8.1 Overview

Language standards should define, and require implementations to provide, means for the user to associate alternate exception handling attributes with blocks (see 4.1 and 7). Alternate exception handlers specify lists of exceptions and actions to be taken for each listed exception if it is signaled. Exception lists might contain:

- one or more of the five exceptions in clause 7: invalid, divideByZero, overflow, underflow, or inexact
- one or more sub-exceptions — sub-cases of the exceptions in clause 7 (*e.g.*, divideByZero); the subexception names are language-defined
- **allExceptions**: all of the exceptions defined in this standard.

All implementations should provide alternate exception handling for allExceptions and any combination of exceptions and subexceptions.

Language standards should provide all the alternate exception handling attributes of this clause. The syntax and scope for such specifications of attribute values are language-defined.

8.2 Resuming alternate exception handling attributes

Resuming alternate exception handling attribute associated with a block means: handle the implied exceptions according to the resuming attribute specified, and resume execution of the associated block. Implementations should support these resuming attributes:

- **default (raiseFlag)**

  Provide the default exception handling (see 7) in the associated block despite alternate exception handling that might be in effect in outer contexts.
- **raiseNoFlag**

  Provide the default exception handling (see 7) without raising the corresponding status flag.
- **mayRaiseFlag**

  Provide the default exception handling (see 7), except languages define whether a flag is raised. Languages may defer to implementations for performance.
- **recordException**

  Provide the default exception handling (see 7) and record the corresponding exception whenever clause 7 recommends raising a flag. Recording an exception means storing a description of the exception, including language-standard-defined details which might include the current operation and operands, and the location of the exception. Language standards define operations to convert to and from character sequences, and to test, save, and restore exception descriptions.
- **substitute(*x*)**

  Specifiable for any exception: replace the default result of such an exceptional operation with a variable or expression *x*. The timing and scope in which *x* is evaluated is language-defined
- **substituteXor(*x*)**

  Specifiable for any exception arising from multiplication or division operations: like substitute(*x*), but replace the default result of such an exceptional operation with |x| and, if |x| is not a NaN, obtaining the sign bit from the XOR of the signs of the operands
- **abruptUnderflow**

  When underflow is signaled because a tiny non-zero result is detected, replace the default result with a zero of the same sign or a minimum normal rounded result of the same sign, raise the underflow flag, and signal inexact.

When roundTiesToEven, roundTiesToAway, or the roundTowardZero attribute is applicable, the rounded result magnitude shall be zero.

When the roundTowardPositive attribute is applicable, the rounded result magnitude shall be the minimum normal magnitude for positive tiny results, and zero for negative tiny results.

When the roundTowardNegative attribute is applicable, the rounded result magnitude shall be the minimum normal magnitude for negative tiny results, and zero for positive tiny results.

This attribute has no effect on the interpretation of subnormal operands.

8.3 Immediate and delayed alternate exception handling attributes

Alternate exception handling associated with a block means: handle the indicated exception according to the attribute specified. If the indicated exception is signaled, then, depending on the exception and the exception handling attribute, the execution of the associated block is either abandoned immediately or continues with default handling. In the latter case the exception handling is delayed and takes place when the associated block terminates normally. Delayed exception handling is fully deterministic, while immediate exception handling might not be deterministic with respect to intermediate results being computed within the associated block..

Language standards should define, and require implementations to provide, these attributes:

- immediate alternate exception handler block associated with a block: if the indicated exception is signaled, abandon execution of the associated block as soon as possible and execute the handler block, then continue execution where execution would have continued after normal termination of the associated block, according to the semantics of the language
- delayed alternate exception handler block associated with a block: if the indicated exception is signaled, handle it by default until the associated block terminates normally, then execute the handler block, then continue execution where execution would have continued after normal termination of the associated block, according to the semantics of the language
- immediate transfer associated with a block: if the indicated exception is signaled, transfer control as soon as possible; no return is possible
- delayed transfer associated with a block: if the indicated exception is signaled, handle it by default until the associated block terminates normally, then transfer control; no return is possible.

A transfer is a language-specific idiom for non-resumable control transfer. Language standards might offer several transfer idioms such as:

- **break:** abandon the associated block and continue execution where execution would continue after normal termination of the associated block, according to the semantics of the language
- **throw exceptionName:** causes an exceptionName not to be handled locally, but rather signaled to the next handling in scope, perhaps the function that invoked the current subprogram, according to the semantics of that language. The invoker might handle exceptionName by default or by alternate handling such as signaling exceptionName to the next higher invoking subprograms
- **goto label:** label might be local or global according to the semantics of the language.

9. Recommended operations

Clause 5 completely specifies operations required for all supported formats available for arithmetic. This clause specifies additional operations recommended for all supported formats available for arithmetic.

9.1 Conforming language- and implementation-defined functions

Language standards and implementations might define floating-point functions, not otherwise defined in this document, that conform to this standard by meeting all the requirements of this subclause. In particular, language standards should define, to be implemented according to this subclause, as many of the functions of 9.2 as is appropriate to the language.

The domain of a function shall be that subset of the affinely extended reals for which the function is well defined, possibly extended by 9.1.1.

A conforming function shall return results correctly rounded for the applicable rounding direction for all operands in its domain, except as modified by 9.1.4. For the purpose of illustration, examples of mathematical functions are given in 9.1.2 which if implemented would encounter cases of special operands.

9.1.1 Exceptions

Except as noted here, functions signal all appropriate exceptions according to 7. All functions shall return a quiet NaN as a result if there is a NaN among a function's operands, except in the cases listed in 9.2.

— invalid: For all functions, signaling NaN operands shall signal invalid.

  Attempts to evaluate a function f outside its domain shall return a quiet NaN and signal invalid, except if the following limit v exists at a representable essential singularity $\boldsymbol{s}$ of f:

  Let f be a function of n variables, and let $\boldsymbol{x} \in \mathrm{R}^n$ be a point in the real domain of f. Let $\boldsymbol{s}$ be the location of the essential singularity. Let $\mu()$ be the measure function for the real domain of f. Let roundTiesToEven() be that function that takes a real number to its nearest representable value in the format of f.

  Then, if there is a value v such that for all open epsilon balls $\mathbf{S}$ containing $\boldsymbol{s}$, where

  $\mu_{\mathrm{notv}}(\boldsymbol{x})$ is $\mu(\{\boldsymbol{x}$ in $\mathbf{S} \mid \mathrm{roundTiesToEven}(f(\boldsymbol{x})) \neq v\})$,

  $\mu_{\mathrm{v}}(\boldsymbol{x})$ is $\mu(\{\boldsymbol{x}$ in $\mathbf{S} \mid \mathrm{roundTiesToEven}(f(\boldsymbol{x})) = v\})$, such that

  $\lim_{\mu(\mathbf{S})\to 0} \mu_{\mathrm{notv}}(\boldsymbol{x}) / \mu_{\mathrm{v}}(\boldsymbol{x}) = 0$

  then v is the computed value of $f(\boldsymbol{s})$ and no exception is signaled. (Thus pow(0, 0) is 1 without exception.)

— divideByZero: A function that has a simple pole for some finite floating-point operand shall return a signed infinity and signal divideByZero. The sign of a simple pole is the sign of its residue if the pole is of even order. It is the sign of its residue times the sign of zero if it is an odd order pole at zero. Otherwise (odd poles other than zero) it is positive for positive x and negative for negative x to preserve sign symmetry even through the evaluation of poles.

— inexact: Functions signal the inexact exception if and only if the numerical result is inexact.

9.1.2 Special operand Zero

Special care shall be taken with $f(0)$ to avoid manifold jumps in the complex plane. Therefore, for any conforming function $f(x)$:

- If $f(x)$ is continuous at zero and its value at zero is some non-zero value c, both $f(-0)$ and $f(+0)$ shall be c; examples include $exp(x)$, $cosh(x)$, $cosPi(x)$, $cos(x)$, $sec(x)$, $sech(x)$
- If $f(x)$ is continuous at zero and its value at zero is zero and the first non-zero term of the Taylor series expansion of $f(x)$ is of the form $c \times x^{2k}$ for some integer k then both $f(-0)$ and $f(+0)$ shall be copySign(0, c); examples include $cosm1(x) = cos(x) - 1$ (for which $f(-0) = f(+0) = -0$), $secm1(x) = sec(x) - 1$ (for which $f(-0) = f(+0) = +0$)
- If $f(x)$ is continuous at zero and its value at zero is zero and the first non-zero term of the Taylor series expansion of $f(x)$ is of the form $c \times x^{2k+1}$ for some integer k then $f(-0)$ shall be copySign(0, $-c$) and $f(+0)$ shall be copySign(0, c); examples include $expm1(x)$, $sinh(x)$, $log1p(x)$, $sinPi(x)$, $atanPi(x)$, $sin(x)$, $tan(x)$, $atan(x)$, $asin(x)$, $tanh(x)$, $asinh(x)$, $atanh(x)$, $Si(x)$ (the sine integral), $erf(x)$ (the error function), $C(x)$ (the Fresnel cosine integral), $S(x)$ (the Fresnel sine integral)
- If $f(x)$ is discontinuous at zero and zero is a domain boundary within the Reals then both $f(-0)$ and $f(+0)$ shall be whichever of $\lim_{x\to 0} -f(x)$ or $\lim_{x\to 0} +f(x)$ is within the domain; examples include $log(x)$, $log2(x)$, $log10(x)$, $asech(x)$, $Ei(x)$ (the exponential integral)
- Except for when f(0) is an essential singularity that meets the conditions defined in subclause 9.2.1, if $f(x)$ is discontinuous at zero but zero is not a domain boundary within the Reals then $f(-0)$ shall be $\lim_{x\to 0} -f(x)$ and $f(+0)$ shall be $\lim_{x\to 0} +f(x)$ and also, if the value of $f(x)$ at zero is not a representable number, that value shall be rounded in a direction consistent with nearby values of the same sign; examples include $acos(x)$ ($\pi/2$ is truncated at zero), $1/x$, $csc(x)$, $cot(x)$, $coth(x)$, $csch(x)$, $Ci(x)$ (the cosine integral), $Gamma(x)$ (the Gamma function), $psi(x)$ (the Digamma function).

9.1.3 Special operand Infinity

If $\lim_{x\to\infty} f(x)$ exists (for either signed infinity) then $f(x)$ takes on that value rounded according to the applicable rounding attribute. If it does not exist (as for, for example, $sin(x)$) then the domain shall be contracted to be contained within the finite numbers and an evaluation of the function at that infinity shall be considered as outside the domain (see 9.1.4).

9.1.4 Domain boundaries

If a domain boundary of a function is a finite number not exactly representable in the operand format and the inverse function takes a manifold jump at that point in the complex plane then when roundTiesToEven or roundTiesToAway is applicable the inverse function shall round its result in order to be contained within the contiguous portion of the forward function's domain. For examples, see the first two cases in 9.2.1.

9.2 Recommended correctly rounded functions

Language standards should define, to be implemented according to , as many of the operations in Table 13 as is appropriate to the language. As with other operations of this standard, the names of the operations in Table 13 do not necessarily correspond to the names that any particular programming language would use.

All functions signal invalid on signaling NaN operands, and inexact on inexact results, as described in 9.1.1; other exceptions as shown in the table.

Table 13—Recommended correctly rounded functions

Operation	Function	Domain	Other exceptions
exp expm1 exp2 exp2m1 exp10 exp10m1 sinh	e^x e^x-1 2^x 2^x-1 10^x 10^x-1 sinh(x)	$[-\infty, +\infty]$	overflow; underflow
log log2 log10	$\log_e(x)$ $\log_2(x)$ $\log_{10}(x)$	$[0, +\infty]$	$x = 0$: divideByZero; $x < 0$: invalid
logp1 log2p1 log10p1	$\log_e(1+x)$ $\log_2(1+x)$ $\log_{10}(1+x)$	$[-1, +\infty]$	$x = -1$: divideByZero; $x < -1$: invalid; underflow
hypot(x, y)	$\sqrt{(x^2+y^2)}$	$[-\infty, +\infty] \times [-\infty, +\infty]$	overflow; underflow
rsqrt	1/sqrt(x)	$[0, +\infty]$	x < 0: invalid x is ±0: divideByZero, default result ±∞
compound(x,n)	$(1+x)^n$	$[-1, +\infty] \times \mathbf{Z}$	x < −1: invalid see also 9.2.1
rootn(x,n)	$x^{1/n}$	$[-\infty, +\infty] \times \mathbf{Z}$	n = 0: invalid x < 0 and n even: invalid n = −1: overflow, underflow
pown(x,n)	x^n	$[-\infty, +\infty] \times \mathbf{Z}$	see 9.2.1
pow(x, y)	x^y	$[-\infty, +\infty] \times [-\infty, +\infty]$	see 9.2.1
sinPi	$\sin(\pi \times x)$	$(-\infty, +\infty)$	$\lvert x\rvert = \infty$: invalid; underflow see also 9.2.1
cosPi	$\cos(\pi \times x)$	$(-\infty, +\infty)$	$\lvert x\rvert = \infty$: invalid see also 9.2.1
atanPi	$\text{atan}(x)/\pi$	$[-\infty, +\infty]$	underflow see also 9.2.1
atan2Pi(y, x)	see text below	$[-\infty, +\infty] \times [-\infty, +\infty]$	underflow
sinh	sinh(x)	$[-\infty, +\infty]$	overflow; underflow
tanh	tanh(x)	$[-\infty, +\infty]$	underflow

Operation	Function	Domain	Other exceptions
cosh	cosh(x)	$[-\infty, +\infty]$	overflow
sin	sin(x)	$(-\infty, +\infty)$	$\|x\| = \infty$: invalid; underflow
tan	tan(x)	$(-\infty, +\infty)$	$\|x\| = \infty$: invalid; underflow
cos	cos(x)	$(-\infty, +\infty)$	$\|x\| = \infty$: invalid
asin	asin(x)	$[-1, +1]$	$\|x\| > 1$: invalid; underflow
acos	acos(x)	$[-1, +1]$	$\|x\| > 1$: invalid
atan	atan(x)	$[-\infty, +\infty]$ for $\|x\| >$ tan(P2), see 9.2.1	underflow
atan2(y, x)	see text below	$[-\infty, +\infty] \times [-\infty, +\infty]$ for $\|$atan2(y, x)$\| >$ P1, see 9.2.1	underflow
acosh	acosh(x)	$[+1, +\infty]$	$x < 1$: invalid
asinh	asinh(x)	$[-\infty, +\infty]$	underflow
atanh	atanh(x)	$[-1, +1]$	underflow

Interval notation is used for the domain: a value adjacent to a bracket is included in the domain and a value adjacent to a parenthesis is not.

9.2.1 Special values

For the hypot function, hypot(±0, ±0) = +0.

For the compound, rootn, and pown functions, n is a finite integral value in logBFormat. When logBFormat is a floating-point format, the behavior of these functions is language-defined when the second operand is non-integral or infinite.

For the compound function:

compound(x, 0) = 1 for any x (even when x is −1 or quiet NaN)
compound(−1, n) = +∞ and signals the divideByZero exception for integral $n < 0$
compound(−1, n) = +0 for even $n > 0$.

For the rootn function:

rootn(±0, n) = ±∞ and signals the divideByZero exception for odd integral $n < 0$
rootn(±0, n) = +∞ and signals the divideByZero exception for even integral $n < 0$
rootn(±0, n) = +0 for even $n > 0$.

For the pown function:

pown(x, 0) = 1 for any x (even a zero or quiet NaN)
pown(±0, n) = ±∞ and signals the divideByZero exception for odd integral $n < 0$
pown(±0, n) = +∞ and signals the divideByZero exception for even integral $n < 0$
pown(±0, n) = +0 for even $n > 0$.

For the pow function:

pow(x, ±0) = 1 for any x (even a zero or quiet NaN)
pow(±0, y) = ±∞ and signals the divideByZero exception for y an odd integer < 0
pow(±0, y) = +∞ and signals the divideByZero exception for $y < 0$ and not an odd integer
pow(±0, y) = +0 for $y > 0$ and not an odd integer

pow(+1, *y*) = 1 for any *y* (even a quiet NaN)
pow(*x*, *y*) returns a quiet NaN and signals the invalid exception for finite $x<0$ and finite non-integer *y*.

For *f* either of sinPi or tanPi, $f(+n)$ is +0 and f(−*n*) is −0 for positive integer *n*. This gives $f(-x) = -f(x)$ for all *x*. cosPi($n + ½$) = +0 for any integer *n*. This gives $f(-x) = f(x)$ for all *x*.

atan2Pi(*y, x*) is the angle, in the range [−1, +1], subtended at the origin by the point (*x, y*) and the positive *x*-axis, which is the argument or phase or imaginary part of the logarithm of the complex number $x+iy$, which is atanPi(*y*/*x*) for $x>0$. atan2Pi(±0, −0) = ±1, atan2Pi(±0, +0) is ±0, atan2Pi(±0, x) is ±1 for $x<0$, atan2Pi(±0, x) is ±0 for $x>0$, atan2Pi(0, 0) does not signal the invalid exception, and atan2Pi(*y*, 0) does not signal the divideByZero exception.

When $|x| >$ tan(P2) for roundTiesToEven or roundTiesToAway, atan(*x*) is copySign(P2, *x*) and might not be correctly rounded (where P2 is π/2 rounded toward zero in the format of *x*). When $|x|>$tan(P2) for directed rounding, atan(x) is correctly rounded to ±P2 or to ±nextUp(P2), in order to support inclusion for interval arithmetic.

atan2(*y, x*) is the angle, in the range [−π, +π], subtended at the origin by the point (*x, y*) and the positive *x*-axis, which is the argument or phase or imaginary part of the logarithm of the complex number $x+iy$, which is atan(*y*/*x*) for $x>0$. atan2(±0, −0) = ±P1, atan2(±0, +0) is ±0, atan2(0, 0) does not signal the invalid exception, and atan2(*y*, 0) does not signal the divideByZero exception.

When $|$atan2(*y, x*)$|>$P1 for roundTiesToEven or roundTiesToAway, then atan2(*y, x*) is copySign(P1, *y*) where P1 is π rounded toward zero in the format of *x*.

Non-standard formats with very large precision relative to exponent range might signal additional exceptions not listed in Table 13. cosPi and log might signal underflow or overflow and tan might signal overflow.

9.3 Operations on dynamic modes for attributes

9.3.1 Operations on individual dynamic modes

Languages standards that define dynamic mode specification for binary or decimal rounding directions shall define corresponding non-computational operations to get and set the applicable value of each specified dynamic mode rounding direction. The applicable value of the rounding direction might have been set by a constant attribute specification or a dynamic-mode assignment, according to the scoping rules of the language. The effect of these operations, if used outside the static scope of a dynamic specification for a rounding direction, is language-defined (and may be unspecified).

Language standards that define dynamic mode specification for binary rounding direction shall define:

— *binaryRoundingDirectionType* **getBinaryRoundingDirection***(void)*
— *void* **setBinaryRoundingDirection**(*binaryRoundingDirectionType*).

Language standards that define dynamic mode specification for decimal rounding direction shall define:

— *decimalRoundingDirectionType* **getDecimalRoundingDirection***(void)*
— *void* **setDecimalRoundingDirection**(*decimalRoundingDirectionType*).

Language standards that define dynamic mode specification for other attributes shall define corresponding operations to get and set those dynamic modes.

9.3.2 Operations on all dynamic modes

Implementations supporting dynamic specification for modes shall provide the following non-computational operations for all dynamic-specifiable modes collectively:

- *modeGroupType* **saveModes(***void*)
  saves the values of all dynamic-specifiable modes as a group
- *void* **restoreModes**(*modeGroupType*)
  restores the values of all dynamic-specifiable modes as a group
- *void* **defaultModes**(*void*)
  sets all dynamic-specifiable modes to default values.

modeGroupType represents the set of dynamically-specifiable modes. The return values of the saveModes operation are for use as operands of the restoreModes operation in the same program; this standard does not require support for any other use.

9.4 Reduction operations

Language standards should define reduction operations, for all supported floating-point formats available for arithmetic, for associative operations like sums, products, sums of products, and products of sums. Unlike the rest of the operations in this standard, these operate on vectors of operands in one format, and implementations may associate in any order, evaluate in any wider format, and short-circuit evaluation when an invalid exception is signaled. Thus numerical results and exceptions signaled might not be identical on different implementations.

In particular, language standards should define the following scaled product reduction operations:

- (*sourceFormat, integralFormat*) **scaledProd** (*source vector, integralFormat*)
  {*pr, sf*} = **scaledProd**(*p, n*) where *p* is a vector of length *n*;
  scaleB(*pr, sf*) is an implementation-defined approximation to $\prod_{(i = 1,n)} p_i$
- (*sourceFormat, integralFormat*) **scaledProdSum** (*source vector, source vector, integralFormat*)
  {*pr, sf*} = **scaledProdSum**(*p, q, n*) where *p* and *q* are vectors of length *n*;
  scaleB(*pr, sf*) is an implementation-defined approximation to $\prod_{(i = 1,n)} (p_i + q_i)$
- (*sourceFormat, integralFormat*) **scaledProdDiff** (*source vector, source vector, integralFormat*)
  {*pr, sf*} = **scaledProdDiff**(*p, q, n*) where *p* and *q* are vectors of length *n*; scaleB(*pr, sf*) is an implementation-defined approximation to $\prod_{(i= 1,n)} (p_i - q_i)$.

These operations attempt to avoid overflow and underflow and compute a scaled product *pr* and a scale factor *sf*. An accurate unscaled product, when *sf* is in the range of *logBFormat*, could be recovered with scaleB(*pr, sf*), in the absence of overflow or underflow.

The vector operands and the scaled product member of the result shall be of the same format. The vector length operand and the scale factor member of the result shall be of the same language-defined format for integral values, *integralFormat*. If *integralFormat* is a floating-point format, it shall have a precision at least as large as *source* and have the same radix.

The implementation of these operations shall use the default exception handling for its internal computations. These operations shall signal the inexact operation and invalid operation exceptions which result from the implementation's use of general computational operations. These operations should avoid signaling overflow and underflow unless the computed scale factor member of the result would exceed the range of *integralFormat*. If implemented with logB, these operations should not signal the divideByZero exception. The return value, when the vector length is less than one, is language-defined.

The preferred exponent for *pr* is 0. If *integralFormat* is a floating-point format, the preferred exponent for *sf* is 0.

10. Expression evaluation

10.1 Expression evaluation rules

Clause 5 of this standard specifies the result of a single arithmetic operation going to an explicit destination. Every operation has an explicit or implicit destination. One rounding occurs to fit the exact result into a destination format. That result is reproducible in that the same operation applied to the same operands under the same attributes produces the same result on all conforming implementations in all languages.

Programming language standards might define syntax for expressions that combine one or more operations of this standard, producing a result to fit an explicit or implicit final destination. When a variable with a declared format is a final destination, as in format conversion to a variable, that declared format of that variable governs its rounding. The format of an implicit destination, or of an explicit destination without a declared format, is defined by language standard expression evaluation rules.

A programming language specifies one or more explicit rules for expression evaluation. A rule for expression evaluation encompasses:

- the order of evaluation of operations
- the formats of implicit intermediate results
- when assignments to explicit destinations round once, and when twice (see below)
- the formats of parameters to generic and non-generic functions
- the formats of results of generic functions.

Languages might permit the programmer to select different language-standard-defined rules for expression evaluation, and might allow implementations to define additional expression evaluation rules and specify the default expression evaluation rule; in these cases language standards should define widenTo attributes as specified below.

Some language standards implicitly convert operands of standard floating-point operations to a common format. Typically, operands are promoted to the widest format of the operands or a widenTo format. However, if the common format is not a superset of the operand formats, then the conversion of an operand to the common format might not preserve the values of the operands. Examples include:

- converting a fixed-point or integer operand to a floating-point format with less precision
- converting a floating-point operand from one radix to another
- converting a floating-point operand to a format with the same radix but with either less range or less precision.

Languages standards should disallow, or provide warnings for, mixed-format operations that would cause implicit conversion that might change operand values.

10.2 Assignments, parameters, and function values

The last operation of many expressions is an assignment to an explicit final destination variable. As a part of expression evaluation rules, language standards shall specify when the next to last operation is performed by rounding at most once to the format of the explicit final destination, and when by rounding as many as two times, first to an implicit intermediate format, and then to the explicit final destination format The latter case does not correspond to any single operation in clause 5 but implies a sequence of two such operations.

In either case, implementations shall never use an assigned-to variable's wider precursor in place of the assigned-to variable's stored value when evaluating subsequent expressions.

When a function has explicitly-declared formal parameter types in scope, the actual parameters shall be rounded if necessary to those explicitly-declared types. When a function does not have explicitly-declared formal parameter types in scope, or is generic, the actual parameters shall be rounded according to language-standard-defined rules.

When a function explicitly declares the type of its return value, the return value shall be rounded to that explicitly-declared type. When the return value type of a function is implicitly defined by language standard rules, the return value shall be rounded to that implicitly-defined type.

10.3 widenTo attributes for expression evaluation

Languages defining generic operations, supporting more than one format available for arithmetic in a particular radix, and defining or allowing more than one way to map expressions in that language into the operations of this standard, should define widenTo attributes for each such format. widenTo attributes are explicitly enabled by the programmer and specify one aspect of expression evaluation: the implicit destination format of language standard-specified generic operations.

In this standard, a computational operation which returns a numeric result first produces an unrounded result as an exact number of infinite precision. That unrounded result is then rounded to a destination format. For certain language standard-specified generic operations, that destination format is implied by the widths of the operands and by the widenTo attribute currently in effect.

An implementation should provide a widenTo attribute for each supported format available for arithmetic.

The following widenTo attributes disable and enable widening of operations in expressions that might be as simple as $z=x+y$ or that might involve several operations on operands of different formats.

- **noWidenTo** attribute: A language standard should define, and require implementations to provide, means for users to specify a noWidenTo attribute, for a block. Destination width is the maximum of the operand widths: generic operations with floating-point operands and results (of the same radix) round results to the widest format among the operands, unless that format is not available for arithmetic, in which case the result should be rounded to the narrowest supported basic format
- **widenToFormat** attributes: A language standard that provides addition, subtraction, multiplication, division, and comparison as generic operators should define, and require implementations to provide, means for users to specify a widenToFormat attribute for each supported format available for arithmetic, for a block. widenToFormat attributes affect the aforementioned operators. Whether and which other generic operators or functions they affect is language-standard-defined. Table 14 lists operators that are suitable for being affected by widenTo attributes. Destination width is the maximum of the width of the widenToFormat and operand widths: affected operations with floating-point operands and results (of the same radix) round results to the widest format among the operands and the widenToFormat. Affected operations (including comparisons and ordering, *e.g.*, maxNum) do not narrow their operands, which might be widened expressions. The widenTo attribute affects the comparison and ordering operations in the same way as arithmetic operations. widenToFormat affects only expressions in the radix of that format.

widenTo attributes do not affect the width of the final rounding to an explicit destination with a declared format, which is always rounded to that format.

widenTo attributes do not affect explicit format conversions within expressions; they are always rounded to the format specified by the conversion.

The widenTo attributes define the width of a generic operation to be the maximum of the widths of its operands and the width of the widenToFormat, if any is in effect. That “maximum” implies an ordering among the formats of the operands whereby one shall be a subset of the other.

Table 14—widenTo operations

Generic operation
destination **addition**(*source1, source2*) *destination* **subtraction**(*source1, source2*) *destination* **multiplication**(*source1, source2*) *destination* **division**(*source1, source2*)
destination **squareRoot** (*source1*)
destination **fusedMultiplyAdd** (*source1, source2, source3*)
destination **minNum**(*source1, source2*) *destination* **maxNum**(*source1, source2*) *destination* **minNumMag**(*source1, source2*) *destination* **maxNumMag**(*source1, source2*)
boolean **compareEqual**(*source1, source2*) *boolean* **compareNotEqual**(*source1, source2*) *boolean* **compareGreater**(*source1, source2*) *boolean* **compareGreaterEqual**(*source1, source2*) *boolean* **compareLess**(*source1, source2*) *boolean* **compareLessEqual**(*source1, source2*) *boolean* **compareSignalingNotGreater**(*source1, source2*) *boolean* **compareSignalingLessUnordered**(*source1, source2*) *boolean* **compareSignalingNotLess**(*source1, source2*) *boolean* **compareSignalingGreaterUnordered**(*source1, source2*) *boolean* **compareQuietGreater**(*source1, source2*) *boolean* **compareQuietGreaterEqual**(*source1, source2*) *boolean* **compareQuietLess**(*source1, source2*) *boolean* **compareQuietLessEqual**(*source1, source2*) *boolean* **compareUnordered**(*source1, source2*) *boolean* **compareQuietNotGreater**(*source1, source2*) *boolean* **compareQuietLessUnordered**(*source1, source2*) *boolean* **compareQuietNotLess**(*source1, source2*) *boolean* **compareQuietGreaterUnordered**(*source1, source2*) *boolean* **compareOrdered**(*source1, source2*)
destination **f**(*source*) or **f(***source1,source2*) for **f** any of the functions in Table 13.

10.4 Value-changing optimizations

A language processor preserves the literal meaning of a floating-point expression if:

- only addition and multiplication can commute the order of their operands
- parentheses and the order of operations are followed unless the programmer licenses associativity
- the distributive law is never applied unless it can be done without altering rounding or exceptions
- the execution of operations is never moved past the boundaries of scope of attribute or dynamic mode specifications.

A language standard should require that execution behavior preserve the literal meaning of the source code and not change the numerical results or exceptions signaled. A language standard should also define, and require implementations to provide, attributes that allow and disallow value-changing optimizations, separately or collectively, for a block. These value-changing optimizations allow more efficient computation of operation results that might differ from the reproducible unoptimized results, but are just as valid for the specific program that explicitly licenses them.

11. Reproducible floating-point results

As described below, reproducible floating-point numerical and status flag results are possible for reproducible operations, with reproducible attributes, operating on reproducible formats, where:

- a reproducible operation is one of the operations described in clause 5 or a supported operation from 9.2 or 9.3
- a reproducible attribute is an attribute that is required in all implementations (see 4)
- a reproducible format is a basic format (see 3).

Programs that can be reliably translated into an explicit or implicit sequence of reproducible operations on reproducible formats produce reproducible results. That is, the same numerical and status flag results are produced.

Reproducible results require cooperation from language standards, language processors, and programmers. A language standard should support reproducible programming. Any conforming language standard supporting reproducible programming shall

- support the reproducible-results attribute
- support a reproducible format by providing all the reproducible operations for that format
- provide means to explicitly or implicitly specify any required sequence of reproducible operations on reproducible formats supported by that language

and shall explicitly define:

- which language feature corresponds to which supported reproducible format
- how to specify in the language each reproducible operation on each supported reproducible format
- one or more unambiguous expression evaluation rules that shall be available for programmer selection on all conforming implementations of that language standard, without deferring any aspect to implementations. If a language standard permits more than one interpretation of a sequence of operations from this standard it shall provide a means of specifying an unambiguous evaluation of that sequence (such as by prescriptive parentheses)
- a reproducible-results attribute, as described in 4.1, with values to indicate when reproducible results are required or reproducible results are not required. Language standards define the default value. When the programmer selects reproducible results required,
  - execution behavior shall preserve the literal meaning (see 10.4) of the source code
  - the invalid exception shall be signaled for fusedMultiplyAdd(0, ∞, *c*) or for fusedMultiplyAdd(∞, 0, *c*) even if *c* is a quiet NaN
  - the underflow exception for binary formats and binary conversion operations shall be signaled if and only if tininess is detected after rounding
  - defined alternate exception handling (see 8) shall be reproducible
  - language processors shall indicate where reproducibility of operations that can affect the results of floating-point operations can not be guaranteed.

Programmers obtain the same floating-point numerical and status flag results on all platforms supporting such a language standard by writing programs

- invoking the reproducible results required attribute
- using only floating-point formats that are reproducible formats
- explicitly, or implicitly via expressions, invoking only reproducible floating-point operations
- invoking only reproducible attributes for rounding, alternate exception handling, and widenTo
- using only integer and non-floating-point formats supported in all implementations of the language standard, and only in ways that avoid signaling implementation-defined integer overflow and divideByZero and other exceptions)
- avoiding value-changing optimizations (see 10.4) and system limits.

Annex A

(informative)

Bibliography

The following documents might be helpful to the reader.

ANSI X3.4–1986, US ASCII Character Set.

IEC 60559:1989, Binary floating-point arithmetic for microprocessor systems (previously designated IEC 559:1989).

ISO/IEC 9899, Second edition 1999–12–01, Programming languages—C.

The Unicode Standard, Version 5.0, The Unicode Consortium, Addison-Wesley Professional, 27 October 2006, ISBN 0-321-48091-0.

S. Boldo and J.-M. Muller, "Some functions computable with a fused-mac", Proceedings of the17th IEEE Symposium on Computer Arithmetic, ISBN 0-7695-2366-8, pp52–28, IEEE Computer Society, 2005.

J.D. Bruguera and T. Lang, "Floating-point Fused Multiply-Add: Reduced Latency for Floating-Point Addition", Proceedings of the 17th IEEE Symposium on Computer Arithmetic, ISBN 0-7695-2366-8, pp42–51, IEEE Computer Society, 2005.

J.T. Coonen, "Contributions to a Proposed Standard for Binary Floating-point Arithmetic", PhD thesis, University of California, Berkeley, 1984.

M.F. Cowlishaw, "Densely-Packed Decimal Encoding", IEE Proceedings — Computers and Digital Techniques, Vol. 149 #3, ISSN 1350-2387, pp102–104, IEE, London, 2002.

M.F. Cowlishaw, "Decimal Floating-Point: Algorism for Computers", Proceedings of the 16th IEEE Symposium on Computer Arithmetic, ISBN 0-7695-1894-X, pp104–111, IEEE Computer Society, 2003.

J.W. Demmel and X. Li. "Faster numerical algorithms via exception handling", IEEE Transactions on Computers, 43(8): pp 983–992, 1994.

F. de Dinechin, A. Ershov, and N. Gast, "Towards the post-ultimate libm", Proceedings of the 17th IEEE Symposium on Computer Arithmetic, ISBN 0-7695-2366-8, pp 288–295, IEEE Computer Society, 2005.

F. de Dinechin, Ch. Q. Lauter, and J.-M. Muller, "Fast and correctly rounded logarithms in double-precision", Theoretical Informatics and Applications, 41, pp. 85–102, EDP Sciences, 2007.

N. Higham, "Accuracy and Stability of Numerical Algorithms", Society for Industrial and Applied Mathematics (SIAM), 1996.

W. Kahan, "Branch Cuts for Complex Elementary Functions, or Much Ado About Nothing's Sign Bit", The State of the Art in Numerical Analysis, (Eds. Iserles and Powell), Clarendon Press, Oxford, 1987.

V. Lefèvre, "New results on the distance between a segment and Z^2. Application to the exact rounding", Proceedings of the 17th IEEE Symposium on Computer Arithmetic, ISBN 0-7695-2366-8, pp68–75, IEEE Computer Society, 2005.

V. Lefèvre and J.-M. Muller, "Worst cases for correct rounding of the elementary functions in double precision", Proceedings of the 15th IEEE Symposium on Computer Arithmetic, ISBN 0-7695-1150-3, pp111–118, IEEE Computer Society, 2001.

P. Markstein, "IA-64 and Elementary Functions: Speed and Precision", ISBN 0-13-018348-2, Prentice Hall, Upper Saddle River, NY, 2000.

R.K. Montoye, E. Hokenek, and S.L. Runyou, "Design of the IBM RISC System/6000 floating-point execution unit", IBM Journal of Research and Development, 34(1), pp59–70, 1990.

J.-M. Muller, “Elementary Functions: Algorithms and Implementation”, second edition, Chapter 10, Birkhaeuser, 2005.

E.M. Schwarz, M.S. Schmookler, and S.D. Trong, “Hardware Implementations of Denormalized Numbers”, Proceedings of the 16th IEEE Symposium on Computer Arithmetic, ISBN 0-7695-1894-X, pp70–78, IEEE Computer Society, 2003.

D. Stehlé, V. Lefèvre, and P. Zimmermann, “Searching worst cases of a one-variable function”, IEEE Transactions on Computers, 54(3), pp 340–346, 2005.

Annex B

(informative)

Program debugging support

B.1 Overview

Implementations of this standard vary in the priorities they assign to characteristics like performance and debuggability (the ability to debug). This annex describes some programming environment features that should be provided by implementations that intend to support maximum debuggability. On some implementations, enabling some of these abilities might be very expensive in performance compared to fully optimized code.

Debugging includes finding the origins of and reasons for numerical sensitivity or exceptions, finding programming errors such as accessing uninitialized storage that are only manifested as incorrect numerical results, and testing candidate fixes for problems that are found.

B.2 Numerical sensitivity

Debuggers should be able to alter the attributes governing handling of rounding or exceptions inside subprograms, even if the source code for those subprograms is not available; dynamic modes might be used for this purpose. For instance, changing the rounding direction or precision during execution might help identify subprograms that are unusually sensitive to rounding, whether due to ill-condition of the problem being solved, instability in the algorithm chosen, or an algorithm designed to work in only one rounding-direction attribute. The ultimate goal is to determine responsibility for numerical misbehavior, especially in separately-compiled subprograms. The chosen means to achieve this ultimate goal is to facilitate the production of small reproducible test cases that elicit unexpected behavior.

B.3 Numerical exceptions

Debuggers should be able to detect, and pause the program being debugged, when a prespecified exception is signaled within a particular subprogram, or within specified subprograms that it calls. To avoid confusion, the pause should happen soon after the event which precipitated the pause. After such a pause, the debugger should be able to continue execution as if the exception had been handled by an alternate handler if specified, or otherwise by the default handler. The pause is associated with an exception and might not be associated with a well-defined source-code statement boundary; insisting on pauses that are precise with respect to the source code might well inhibit optimization.

Debuggers should be able to raise and lower status flags.

Debuggers should be able to examine all the status flags left standing at the end of a subprogram's or whole program's execution. These capabilities should be enhanced by implementing each status flag as a reference to a detailed record of its origin and history. By default, even a subprogram presumed to be debugged should at least insert a reference to its name in an status flag and in the payload of any new quiet NaN produced as a floating-point result of an invalid operation. These references indicate the origin of the exception or NaN.

Debuggers should be able to maintain tables of histories of quiet NaNs, using the NaN payload to index the tables.

Debuggers should be able to pause at every floating-point operation, without disrupting a program's logic for dealing with exceptions. Debuggers should display source code lines corresponding to machine instructions whenever possible.

For various purposes a signaling NaN could be used as a reference to a record containing a numerical value extended by an exception history, wider exponent, or wider significand. Consequently debuggers should be

able to cause bitwise operations like negate, abs, and copySign, which are normally silent, to detect signaling NaNs. Furthermore the signaling attribute of signaling NaNs should be able to be enabled or disabled globally or within a particular context, without disrupting or being affected by a program's logic for default or alternate handling of other invalid exceptions..

5 B.4 Programming errors

Debuggers should be able to define some or all NaNs as signaling NaNs that signal an exception every time they are used. In formats with superfluous bit patterns not generated by arithmetic, such as non-canonical significand fields in decimal formats, debuggers should be able to enable signaling-NaN behavior for data containing such bit patterns.

10 Debuggers should be able to set uninitialized storage and variables, such as heap and common space to specific bit patterns such as all-zeros or all-ones which are helpful for finding inadvertent usages of such variables; those usages might prove refractory to static analysis if they involve multiple aliases to the same physical storage.

More helpful, and requiring correspondingly more software coordination to implement, are debugging
15 environments in which all floating-point variables, including automatic variables each time they are allocated on a stack, are initialized to signaling NaNs that reference symbol table entries describing their origin in terms of the source program. Such initialization would be especially useful in an environment in which the debugger is able to pause execution when a prespecified exception is signaled or flag is raised.