SLOVENSKI STANDARD
SIST HD 592 S1:1997

01-avgust-1997

Binarna aritmetika s plavajočo vejico za mikroprocesorske sisteme (IEC 60559:1989)

Binary floating-point arithmetic for microprocessor systems

Binäre Gleitpunkt-Arithmetik für Mikroprozessor-Systeme

Arithmétique binaire en virgule flottante pour systèmes à microprocesseurs

Ta slovenski standard je istoveten z: **HD 592 S1:1991**

ICS:

35.160 Mikroprocesorski sistemi Microprocessor systems

SIST HD 592 S1:1997 **en**

HARMONIZATION DOCUMENT HD 592 S1

DOCUMENT D'HARMONISATION

HARMONISIERUNGSDOKUMENT May 1991

UDC 621.382.049.77.037.372

Descriptors: Binary floating point arithmetic, microprocessor systems

ENGLISH VERSION

BINARY FLOATING-POINT ARITHMETIC FOR MICROPROCESSOR SYSTEMS (IEC 559:1989)

Arithmétique binaire en virgule flottante pour systèmes à microprocesseurs (CEI 559:1989)

Binäre Gleitpunkt-Arithmetik für Mikroprozessor-Systeme (IEC 559:1989)

This Harmonization Document was approved by CENELEC on 1991-03-15. CENELEC members are bound to comply with the CEN/CENELEC Internal Regulations which stipulate the conditions for implementation of this Harmonization Document on a national level.

Up-to-date lists and bibliographical references concerning national implementation may be obtained on application to the Central Secretariat or to any CENELEC member.

This Harmonization Document exists in three official versions (English, French, German).

CENELEC members are the national electrotechnical committees of Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and United Kingdom.

CENELEC

European Committee for Electrotechnical Standardization
Comité Européen de Normalisation Electrotechnique
Europäisches Komitee für Elektrotechnische Normung

Central Secretariat: rue de Stassart 35, B-1050 Brussels

Ref. No. HD 592 S1:1991 E

FOREWORD

The CENELEC questionnaire procedure, performed for finding out whether or not the International Standard IEC 559:1989 could be accepted without textual changes, has shown that no CENELEC common modifications were necessary for the acceptance as Harmonization Document.
The reference document was submitted to the CENELEC members for formal vote and was approved by CENELEC as HD 592 S1 on 15 March 1991.

The following dates were fixed:

- latest date of announcement of the HD at national level	(doa)	1991-09-01
- latest date of publication of a harmonized national standard	(dop)	1992-03-01
- latest date of withdrawal of conflicting national standards	(dow)	1992-03-01

ENDORSEMENT NOTICE

The text of the International Standard IEC 559:1989 was approved by CENELEC as a Harmonization Document without any modification.

NORME INTERNATIONALE

INTERNATIONAL STANDARD

CEI
IEC
60559

Deuxième édition
Second edition
1989-01

Arithmétique binaire en virgule flottante pour systèmes à microprocesseurs

Binary floating-point arithmetic for microprocessor systems

International Electrotechnical Commission 3, rue de Varembé Geneva, Switzerland
Telefax: +41 22 919 0300 e-mail: inmail@iec.ch IEC web site http: //www.iec.ch

IEC Commission Electrotechnique Internationale
International Electrotechnical Commission
Международная Электротехническая Комиссия

CODE PRIX
PRICE CODE S

Pour prix, voir catalogue en vigueur
For price, see current catalogue

CONTENTS

INTERNATIONAL ELECTROTECHNICAL COMMISSION

BINARY FLOATING-POINT ARITHMETIC FOR MICROPROCESSOR SYSTEMS

FOREWORD

1) The formal decisions or agreements of the IEC on technical matters, prepared by Technical Committees on which all the National Committees having a special interest therein are represented, express, as nearly as possible, an international consensus of opinion on the subjects dealt with.

2) They have the form of recommendations for international use and they are accepted by the National Committees in that sense.

3) In order to promote international unification, the IEC expresses the wish that all National Committees should adopt the text of the IEC recommendation for their national rules in so far as national conditions will permit. Any divergence between the IEC recommendation and the corresponding national rules should, as far as possible, be clearly indicated in the latter.

PREFACE

This standard has been prepared by Sub-Committee 47B: Microprocessor systems, of IEC Technical Committee No. 47: Semiconductor devices. (This Sub-Committee has been taken over by ISO/IEC JTC 1.)

This second edition of IEC Publication 559 replaces the first edition issued in 1982.

The text of this standard is based on the following documents:

Six Months' Rule	Report on Voting
47B(CO)19	47B(CO)26

Full information on the voting for the approval of this standard can be found in the Voting Report indicated in the above table.

BINARY FLOATING-POINT ARITHMETIC FOR MICROPROCESSOR SYSTEMS

1. Scope

1.1 *Implementation objectives*

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. It is the environment that the programmer or user of the system sees that conforms or fails to conform to this standard. Hardware components that require software support to conform shall not be said to conform apart from such software.

1.2 *Inclusions*

This standard specifies:

1) basic and extended floating-point number formats;
2) add, subtract, multiply, divide, square root, remainder and compare operations;
3) conversions between integer and floating-point numbers;
4) conversions between different floating-point formats;
5) conversions between basic format floating-point numbers and decimal strings, and
6) floating-point exceptions and their handling, including non-numbers (NaNs).

1.3 *Exclusions*

This standard does not specify:

1) formats of decimal strings and integers;
2) interpretation of the signs and significant fields of NaNs, or
3) binary ↔ decimal conversions to and from extended formats.

2. Definitions

Biased exponent

The sum of the exponent and a constant (bias) chosen to make the biased exponent's range non-negative.

Binary floating-point number

A bit-string characterized by three components: a sign, a signed exponent, and a significand. Its numerical value, if any, is the signed product of its significand and two raised to the power of its exponent. In this standard a bit-string is not always distinguished from a number it may represent.

Denormalized number

A nonzero floating-point number, the exponent of which has a reserved value, usually the format's minimum, and the explicit or implicit leading significant bit of which is zero.

Destination

The location for the result of a binary or unary operation. The destination may be either explicitly designated by the user or implicitly supplied by the system (e.g. intermediate results in sub-expressions 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 as well as the operands' values.

Exponent

The component of a binary floating-point number that normally signifies the integer power to which two is raised in determining the value of the represented number. Occasionally the exponent is called the signed or unbiased exponent.

Fraction

The field of the significand that lies to the right of its implied binary point.

Mode

A variable that a user may set, sense, save and restore, to control the execution of subsequent arithmetic operations. The default mode is the mode that a program can assume to be in effect unless an explicitly contrary statement is included either in the program or in its specification.

The following modes shall be implemented:

1) rounding, to control the direction of rounding errors, and in certain implementations.

2) rounding precision, to shorten the precision of results. The implementor may, at his option, implement the following modes:

3) traps disabled/enabled, to handle exceptions.

NaN

Not a number; a symbolic entity encoded in floating-point format. There are two types of NaNs (see 6.2). Signalling NaNs signal the invalid operation exception (see 7.1) whenever they appear as operands. Quiet NaNs propagate through almost every arithmetic operation without signalling exceptions.

Result

The bit-string (usually representing a number) that is delivered to the destination.

Significant

The component of a binary floating-point number which consists of an explicit or implicit leading bit to the left of its implied binary point and a fraction field to the right.

Shall

The word "shall" signifies that which is obligatory in any conforming implementation.

Should

The word "should" signifies that which is strongly recommended as being in keeping with the intent of the standard, although architectural or other constraints beyond the scope of this standard may, on occasion, render the recommendations impractical.

Status flag

A variable that may take two states, set and clear. A user may clear a flag, copy it, or restore it to a previous state. When set, a status flag may contain additional system-dependent information, possibly inaccessible to some users. The operations of this standard may, as a side-effect, set some of the following flags: inexact result, underflow, overflow, divide by zero and invalid operation.

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.

3. Formats

This standard defines four floating-point formats in two groups, basic and extended, each having two widths, single and double. The standard levels of implementation are distinguished by the combinations of formats supported.

3.1 *Sets of values*

This sub-clause concerns only the numerical values representable within a format, not the encodings which are the subject of the following sub-clauses. The only values representable in a chosen format are those specified via the following three integer parameters:

P = number of significant bits (precision)

E_{max} = maximum exponent, and

E_{min} = minimum exponent

Each format's parameters are displayed in Table 1. Within each format just the following entities shall be provided:

Numbers of the form $(-1)^s 2^E (b_0 b_1 b_2 \ldots b_{p-1})$

where:

s is 0 or 1;
E is any integer between E_{min} and E_{max} inclusive, and each b_i is 0 or 1.

Two infinities, $+\infty$ and $-\infty$;
at least one signalling NaN, and
at least one quiet NaN.

Table 1 - Summary of format parameters

Parameter	Format			
	Single	Single Extended	Double	Double Extended
P	24	≥32	53	≥64
E_{max}	+127	≥+1 023	+1 023	≥+16 383
E_{min}	-126	≤-1 022	-1 022	≤-16 382
Exponent bias	+127	Unspecified	+1 023	Unspecified
Exponent width (bits)	8	≥11	11	≥15
Format width (bits)	32	≥43	64	≥79

The foregoing description enumerates some values redundantly, for example:

$$2^0(1.0) = 2^1(0.1) = 2^2(0.01) = \ldots$$

However, the encodings of such nonzero values may be redundant only in extended formats (see 3.3). The nonzero values of the form $\pm 2^{E_{min}} (0 \cdot b_1 b_2 \ldots b_{p-1})$ are called denormalized. Reserved exponents