

**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**

© IEC 1989 Droits de reproduction réservés — Copyright - all rights reserved

Aucune partie de cette publication ne peut être reproduite ni utilisée sous quelque forme que ce soit et par aucun procédé, électronique ou mécanique, y compris la photocopie et les microfilms, sans l'accord écrit de l'éditeur.

No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the publisher.

International Electrotechnical Commission
Telefax: +41 22 919 0300

e-mail: inmail@iec.ch

3, rue de Varembé Geneva, Switzerland
IEC web site <http://www.iec.ch>



Commission Electrotechnique Internationale
International Electrotechnical Commission
Международная Электротехническая Комиссия

CODE PRIX
PRICE CODE

S

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

SOMMAIRE

	Pages
PREAMBULE	4
PREFACE	4
Articles	
1. Domaine d'application	6
1.1 Objectifs de réalisation	6
1.2 Inclusions	6
1.3 Exclusions	6
2. Définitions	6
3. Formats	10
3.1 Ensembles de valeurs	12
3.2 Formats de base	14
3.3 Formats étendus	16
3.4 Combinaisons de formats	16
4. Arrondi	18
4.1 Arrondi au plus près	18
4.2 Arrondis orientés	18
4.3 Précision d'arrondi	18
5. Opérations	20
5.1 Arithmétique	20
5.2 Racine carrée	22
5.3 Conversions des formats virgule flottante	22
5.4 Conversion entre virgule flottante et entier	22
5.5 Arrondi de nombres en virgule flottante vers une valeur entière	22
5.6 Conversion binaire-décimale	22
5.7 Comparaison	26
6. Infini, non-nombres et zéro signé	30
6.1 Arithmétique de l'infini	30
6.2 Opérations avec des non-nombres	30
6.3 Bit de signe	32
7. Exceptions	32
7.1 Opérations invalides	32
7.2 Division par zéro	34
7.3 Dépassement de capacité	34
7.4 Dépassement de capacité inférieur	36
7.5 Inexactitude	38
8. Déroutements	38
8.1 Routine de traitement de déroutement	40
8.2 Précédence	40
ANNEXE A - Fonctions et prédicats recommandés	42



CONTENTS

	Page
FOREWORD	5
PREFACE	5
Clause	
1. Scope	7
1.1 Implementation objectives	7
1.2 Inclusions	7
1.3 Exclusions	7
2. Definitions	7
3. Formats	11
3.1 Sets of values	13
3.2 Basic formats	15
3.3 Extended formats	17
3.4 Combinations of formats	17
4. Rounding	19
4.1 Round to nearest	19
4.2 Directed roundings	19
4.3 Rounding precision	19
5. Operations	21
5.1 Arithmetic	21
5.2 Square root	23
5.3 Floating-point format conversions	23
5.4 Conversions between floating-point and integer	23
5.5 Round floating-point number to integral value	23
5.6 Binary ↔ decimal conversion	23
5.7 Comparison	27
6. Infinity, NaNs and signed zero	31
6.1 Infinity arithmetic	31
6.2 Operations with NaNs	31
6.3 The sign bit	33
7. Exceptions	33
7.1 Invalid operations	33
7.2 Division by zero	35
7.3 Overflow	35
7.4 Underflow	37
7.5 Inexact	39
8. Traps	39
8.1 Trap handler	41
8.2 Precedence	41
APPENDIX A - Recommended functions and predicates	43

COMMISSION ELECTROTECHNIQUE INTERNATIONALE

ARITHMETIQUE BINAIRE EN VIRGULE FLOTTANTE
POUR SYSTEMES A MICROPROCESSEURS

PREAMBULE

- 1) Les décisions ou accords officiels de la CEI en ce qui concerne les questions techniques, préparés par des Comités d'Etudes où sont représentés tous les Comités nationaux s'intéressant à ces questions, expriment dans la plus grande mesure possible un accord international sur les sujets examinés.
- 2) Ces décisions constituent des recommandations internationales et sont agréées comme telles par les Comités nationaux.
- 3) Dans le but d'encourager l'unification internationale, la CEI exprime le voeu que tous les Comités nationaux adoptent dans leurs règles nationales le texte de la recommandation de la CEI, dans la mesure où les conditions nationales le permettent. Toute divergence entre la recommandation de la CEI et la règle nationale correspondante doit, dans la mesure du possible, être indiquée en termes clairs dans cette dernière.

PREFACE

La présente norme a été établie par le Sous-Comité 47B: Systèmes à microprocesseurs, du Comité d'Etudes n° 47 de la CEI: Dispositifs à semiconducteurs. (Ce Sous-Comité a été repris par l'ISO/IEC JTC 1.)

Cette deuxième édition de la Publication 559 remplace la première édition parue en 1982.

Le texte de cette norme est issu des documents suivants:

Règle des Six Mois	Rapport de vote
47B(BC)19	47B(BC)26

Le rapport de vote indiqué dans le tableau ci-dessus donne toute information sur le vote ayant abouti à l'approbation de cette norme.

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(C0)19	47B(C0)26

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

ARITHMETIQUE BINAIRE EN VIRGULE FLOTTANTE POUR SYSTEMES A MICROPROCESSEURS

1. Domaine d'application

1.1 *Objectifs de réalisation*

L'objectif est qu'une réalisation d'un système à virgule flottante conforme à la présente norme puisse être effectuée entièrement par logiciel, entièrement par matériel, ou par une combinaison quelconque de logiciel et de matériel. C'est l'environnement que le programmeur ou l'utilisateur voit qui est conforme ou non conforme à cette norme. Les composants matériels qui nécessitent un support logiciel pour devenir conformes ne doivent pas être qualifiés de conformes indépendamment d'un tel logiciel.

1.2 *Inclusions*

Cette norme spécifie:

- 1) les formats de base et étendu des nombres en virgule flottante;
- 2) les opérations d'addition, de soustraction, de multiplication, de division, de calcul d'une racine carrée, du calcul d'un reste et de comparaison;
- 3) les conversions entre nombres entiers et nombres en virgule flottante;
- 4) les conversions entre différents formats en virgule flottante;
- 5) les conversions entre les nombres en virgule flottante en format de base et les chaînes décimales, et
- 6) la détection et le traitement des conditions d'exception pour les nombres en virgule flottante y compris les non-nombres ("NaN").

1.3 *Exclusions*

Cette norme ne spécifie pas:

- 1) les formats des chaînes décimales et des entiers;
- 2) l'interprétation des champs de signe et de mantisse des non-nombres ("NaN"), ou
- 3) les conversions de binaire à décimal et réciproquement pour les formats étendus.

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.