

COMMISSION ÉLECTROTECHNIQUE INTERNATIONALE
NORME DE LA CEI

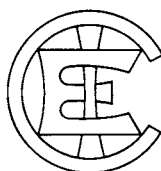
INTERNATIONAL ELECTROTECHNICAL COMMISSION
IEC STANDARD

Publication 559

Première édition — First edition
1982

**Arithmétique binaire en virgule flottante
pour systèmes à microprocesseurs**

**Binary floating point arithmetic
for microprocessor systems**



© CEI 1982

Droits de reproduction réservés — Copyright - all rights reserved

Bureau Central de la Commission Electrotechnique Internationale

3, rue de Varembe
Genève, Suisse

Révision de la présente publication

Le contenu technique des publications de la CEI est constamment revu par la Commission afin d'assurer qu'il reflète bien l'état actuel de la technique.

Les renseignements relatifs à ce travail de révision, à l'établissement des éditions révisées et aux mises à jour peuvent être obtenus auprès des Comités nationaux de la CEI et en consultant les documents ci-dessous:

- **Bulletin de la CEI**
- **Annuaire de la CEI**
- **Catalogue des publications de la CEI**
Publié annuellement

Terminologie

En ce qui concerne la terminologie générale, le lecteur se reportera à la Publication 50 de la CEI: Vocabulaire Electrotechnique International (V.E.I.), qui est établie sous forme de chapitres séparés traitant chacun d'un sujet défini, l'Index général étant publié séparément. Des détails complets sur le V.E.I. peuvent être obtenus sur demande.

Les termes et définitions figurant dans la présente publication ont été soit repris du V.E.I., soit spécifiquement approuvés aux fins de cette publication.

Symboles graphiques et littéraux

Pour les symboles graphiques, symboles littéraux et signes d'usage général approuvés par la CEI, le lecteur consultera:

- la Publication 27 de la CEI: Symboles littéraux à utiliser en électrotechnique;
- la Publication 117 de la CEI: Symboles graphiques recommandés.

Les symboles et signes contenus dans la présente publication ont été soit repris des Publications 27 ou 117 de la CEI, soit spécifiquement approuvés aux fins de cette publication.

Revision of this publication

The technical content of IEC publications is kept under constant review by the IEC, thus ensuring that the content reflects current technology.

Information on the work of revision, the issue of revised editions and amendment sheets may be obtained from IEC National Committees and from the following IEC sources:

- **IEC Bulletin**
- **IEC Yearbook**
- **Catalogue of IEC Publications**
Published yearly

Terminology

For general terminology, readers are referred to IEC Publication 50: International Electrotechnical Vocabulary (I.E.V.), which is issued in the form of separate chapters each dealing with a specific field, the General Index being published as a separate booklet. Full details of the I.E.V. will be supplied on request.

The terms and definitions contained in the present publication have either been taken from the I.E.V. or have been specifically approved for the purpose of this publication.

Graphical and letter symbols

For graphical symbols, and letter symbols and signs approved by the IEC for general use, readers are referred to:

- IEC Publication 27: Letter symbols to be used in electrical technology;
- IEC Publication 117: Recommended graphical symbols.

The symbols and signs contained in the present publication have either been taken from IEC Publications 27 or 117, or have been specifically approved for the purpose of this publication.

COMMISSION ÉLECTROTECHNIQUE INTERNATIONALE
NORME DE LA CEI

INTERNATIONAL ELECTROTECHNICAL COMMISSION
IEC STANDARD

Publication 559

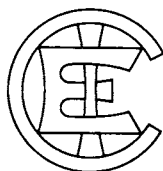
Première édition — First edition
1982

**Arithmétique binaire en virgule flottante
pour systèmes à microprocesseurs**

**Binary floating point arithmetic
for microprocessor systems**

Mots clés: microprocesseurs;
opérations arithmétiques binaires;
exigences; propriétés;
définitions.

Key words: microprocessors;
binary arithmetical operations;
requirements; properties;
definitions.



© CEI 1982

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.

Bureau Central de la Commission Electrotechnique Internationale

3, rue de Varembe
Genève, Suisse

SOMMAIRE

	Pages
PRÉAMBULE	4
PRÉFACE	4
Articles	
1. Domaine d'application	6
2. Objet	6
3. Terminologie	8
4. Formats	12
5. Arrondi	16
6. Opérations	18
7. Infini, NaN et zéro signé	24
8. Arithmétique non normalisée et dénormalisée	26
9. Anomalies	28
10. Pièges	32
ANNEXE — Fonctions et prédicats recommandés	36

Withholding

CONTENTS

	Page
FOREWORD	5
PREFACE	5
Clause	
1. Scope	7
2. Object	7
3. Terminology	9
4. Formats	13
5. Rounding	17
6. Operations	19
7. Infinity, NaNs and signed zero	25
8. Unnormalized and denormalized arithmetic	27
9. Exceptions	29
10. Traps	33
APPENDIX — Recommended functions and predicates	37

Witholdano

COMMISSION ÉLECTROTECHNIQUE INTERNATIONALE

ARITHMÉTIQUE BINAIRE EN VIRGULE FLOTTANTE
POUR SYSTÈMES À MICROPROCESSEURS

PRÉAMBULE

- 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 vœu 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.

PRÉFACE

La présente norme a été préparée par le Sous-Comité 47B: Systèmes à microprocesseurs, du Comité d'Etudes N° 47 de la CEI: Dispositifs à semiconducteurs et circuits intégrés.

Un premier projet* fut discuté lors de la réunion tenue à Montreux en 1981. A la suite de cette réunion, un projet, document 47B(Bureau Central)2, fut soumis à l'approbation des Comités nationaux suivant la Règle des Six Mois en novembre 1981.

Les Comités nationaux des pays suivants se sont prononcés explicitement en faveur de la publication:

Afrique du Sud (République d')
Allemagne
Autriche
Belgique
Canada
Egypte
Espagne
Etats-Unis d'Amérique
France
Irlande

Italie
Japon
Pays-Bas
Pologne
Roumanie
Suède
Suisse
Union des Républiques
Socialistes Soviétiques
Yougoslavie

* Ce premier projet fut préparé d'après le projet P754, version 8.0, du I.E.E.E.

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 and Integrated Circuits.

A first draft* was discussed at the meeting held in Montreux in 1981. As a result of this meeting, a draft, Document 47B(Central Office)2, was submitted to the National Committees for approval under the Six Months' Rule in November 1981.

The National Committees of the following countries voted explicitly in favour of publication:

Austria	Poland
Belgium	Romania
Canada	South Africa (Republic of)
Egypt	Spain
France	Sweden
Germany	Switzerland
Ireland	Union of Soviet
Italy	Socialist Republics
Japan	United States of America
Netherlands	Yugoslavia

* This first draft was prepared in accordance with draft 8.0 of I.E.E.E. Task P754.

COMMISSION ÉLECTROTECHNIQUE INTERNATIONALE

**ARITHMÉTIQUE BINAIRE EN VIRGULE FLOTTANTE
POUR SYSTÈMES A MICROPROCESSEURS****1. Domaine d'application**

Cette norme est applicable à l'arithmétique binaire en virgule flottante pour systèmes à microprocesseurs.

Elle précise:

- a) les formats des nombres en virgule flottante;
- b) les résultats pour l'addition, la soustraction, la multiplication, la division, la racine carrée, le calcul d'un reste et la comparaison;
- c) les conversions entre entiers et nombres en virgule flottante;
- d) les conversions entre différents formats en virgule flottante;
- e) les conversions entre le format de base (voir le paragraphe 4.1) des nombres en virgule flottante et les chaînes décimales;
- f) la détection et le traitement des anomalies des nombres en virgule flottante, y compris les non-nombres («NaN»).

Cette norme ne spécifie pas:

- a) la représentation des nombres entiers;
- b) l'interprétation des signes et champs fractionnaires des non-nombres («NaN»);
- c) les conversions de binaire en décimal et inverse avec les formats étendus;
- d) les formats des chaînes décimales.

2. Objet

L'objet de cette norme est de définir de quelle façon les nouveaux systèmes doivent manipuler l'arithmétique binaire en virgule flottante. Il est prévu que les réalisations d'un système de calcul en virgule flottante conforme à cette norme puissent se faire entièrement en logiciel, entièrement en matériel ou dans une combinaison quelconque des deux.

INTERNATIONAL ELECTROTECHNICAL COMMISSION

**BINARY FLOATING POINT ARITHMETIC
FOR MICROPROCESSOR SYSTEMS****1. Scope**

This standard applies to binary floating-point arithmetic used in microprocessor systems.

This standard specifies:

- a) floating-point number formats;
- b) the results for add, subtract, multiply, divide, square root, remainder and compare;
- c) conversions between integers and floating-point numbers;
- d) conversions between different floating-point formats;
- e) conversions between basic format (see Sub-clause 4.1) floating-point numbers and decimal strings;
- f) floating-point exceptions and their handling, including non-numbers (NaNs).

This standard does not specify:

- a) integer representation;
- b) the interpretation of the signs and the fraction field of non-numbers (NaNs);
- c) binary - decimal conversions to and from extended formats;
- d) formats of decimal strings.

2. Object

The object of this standard is to define ways for new systems to perform binary floating-point arithmetic. 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 hardware and software.

C'est tout l'environnement réel que le programmeur ou l'utilisateur du système voit qui est conforme ou non à cette norme. Les composants matériels qui nécessitent un support logiciel spécial pour être conformes à la norme ne doivent pas être dits conformes sans mention de ce logiciel spécial.

3. Terminologie

3.1 Niveau d'exigence

Dans cette norme, l'usage du verbe:

- «doit» implique que le respect de la spécification est exigé pour être compatible avec la norme;
- «devrait» implique que le respect de la spécification est fortement recommandé, mais n'est pas obligatoire pour être compatible avec la norme.

3.2 Définitions

3.2.1 Utilisateur

Est considéré comme utilisateur d'un système en virgule flottante toute personne, tout matériel ou programme, non précisés dans cette norme, qui a accès aux opérations effectuées dans l'environnement logiciel spécifié dans cette norme et qui les contrôle.

3.2.2 Nombre binaire exprimé en virgule flottante

Chaîne de bits caractérisée par trois éléments: un signe, un exposant signé et une mantisse. Sa valeur numérique, si elle existe, est le produit signé de sa mantisse par deux, à la puissance donnée par l'exposant. Dans cette norme, une chaîne de bits n'est pas toujours distinguée du nombre qu'elle peut représenter.

3.2.3 Exposant

Cet élément d'un nombre binaire exprimé en virgule flottante indique généralement la puissance à laquelle deux est élevé pour calculer la valeur du nombre représenté. Occasionnellement, l'exposant est appelé «exposant signé» ou «sans excédent».

3.2.4 Exposant avec excédent

Somme de l'exposant et d'une constante (excédent ou biais) choisie de sorte que le champ de l'exposant avec excédent ne soit jamais négatif.

3.2.5 Mantisse

Élément d'un nombre binaire exprimé en virgule flottante constitué d'un bit significatif explicite ou implicite à gauche de la virgule et d'un champ fractionnaire à droite de la virgule.

3.2.6 Partie fractionnaire

Champ de la mantisse situé à droite de la virgule correspondante.

It is the actual environment which 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.

3. Terminology

3.1 Degree of requirements

In this standard, the auxiliary verb:

- “shall” implies that compliance with a requirement is mandatory for compliance with the standard;
- “should” implies that compliance with a requirement is strongly recommended but is not mandatory for compliance with the standard.

3.2 Definitions

3.2.1 User

The user of a floating-point system is considered to be 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.

3.2.2 Binary floating-point number

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

3.2.3 Exponent

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

3.2.4 Biased exponent

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

3.2.5 Mantissa

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

3.2.6 Fraction

That field of the mantissa that lies to the right of its implied binary point.

3.2.7 Zéro normalisé

L'exposant a la valeur minimale correspondant au format et la mantisse est zéro. Le zéro normalisé peut avoir soit un signe positif soit un signe négatif. Seuls les formats étendus ont des zéros non normalisés (voir paragraphe 3.2.9).

3.2.8 Nombre dénormalisé

L'exposant a la valeur minimale correspondant au format, le bit de poids fort de la mantisse (explicite ou implicite) est nul et le nombre n'est pas un zéro normalisé. Dénormaliser un nombre binaire exprimé en virgule flottante signifie décaler sa mantisse à droite tout en incrémentant son exposant, jusqu'à ce qu'on obtienne un nombre dénormalisé.

3.2.9 Nombre non normalisé

Ce type de nombre n'apparaît que pour le format étendu. L'exposant du nombre est supérieur à la valeur minimale correspondant au format et le bit le plus significatif explicite est nul. Si la mantisse est nulle, c'est un zéro non normalisé.

3.2.10 Normaliser

Si le nombre est non nul, normaliser revient à décaler sa mantisse à gauche en décrémentant son exposant jusqu'à ce que le bit de poids fort de la mantisse soit un; l'exposant est traité comme si son champ était illimité. Si la mantisse est nulle, le nombre est remplacé par le zéro normalisé. La normalisation d'un nombre ne change pas son signe.

3.2.11 Non-nombre («NaN»)

Entité symbolique codée selon le format virgule flottante. Voir l'article 4 et le paragraphe 7.2.

3.2.12 Indicateur d'état

Variable qui peut prendre deux états: actif (à 1) ou inactif (à 0). Un programme peut désactiver ou copier un indicateur. Un indicateur d'état actif peut contenir des informations additionnelles qui dépendent du système et qui peuvent être inaccessibles à l'utilisateur. Les opérations définies par cette norme peuvent avoir comme effet secondaire d'activer certains indicateurs: résultat inexact, dépassement de capacité inférieure, dépassement de capacité (supérieure), division par zéro et opération invalide.

3.2.13 Résultat

Chaque opération unaire ou binaire mémorise son résultat dans un emplacement de résultat, qui peut être explicitement désigné par l'utilisateur ou fourni implicitement par le système (par exemple résultats intermédiaires dans des sous-expressions ou arguments de procédures). Quelques langages placent les résultats intermédiaires dans des emplacements de résultats qui ne sont pas sous le contrôle de l'utilisateur. Néanmoins cette norme définit le résultat d'une opération en donnant le format de l'emplacement du résultat aussi bien que les valeurs des opérandes.

3.2.14 Modes

Variable qu'un programme peut activer, tester, sauver et rétablir afin de commander l'exécution des opérations arithmétiques concernées. Le mode par défaut est le mode

3.2.7 *Normal zero*

The exponent is the minimum value of the format, and the mantissa is zero. Normal zero may have either a positive or a negative sign. Only the extended formats have any unnormalized zeros (see Sub-clause 3.2.9).

3.2.8 *Denormalized Number*

The exponent of the number is the minimum value of the format, the explicit or implicit leading bit of the significand is zero, and the number is not normal zero. To denormalize a binary floating-point number means to shift its mantissa right while incrementing its exponent, until it becomes a denormalized number.

3.2.9 *Unnormalized Number*

This type of number occurs only in the extended format. The exponent of the number is greater than the minimum value of the format, and the explicit leading bit is zero. If the mantissa is zero, this is an unnormalized zero.

3.2.10 *Normalize*

If the number is nonzero, the number is normalized by shifting its mantissa left while decrementing its exponent until the leading bit of the mantissa becomes one; the exponent is regarded as if its range were unlimited. If the mantissa is zero, the number is replaced by a normal zero. Normalizing a number does not change its sign.

3.2.11 *Non-Number (NaN)*

This is a symbolic entity encoded in floating-point format. See Clause 4 and Sub-clause 7.2.

3.2.12 *Status flag*

A variable which may take two states, set and clear. A program may clear or copy a flag. When set, a status flag may contain additional system-dependent information, possibly inaccessible to the user. 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.

3.2.13 *Destination*

Every unary or binary operation delivers its result to a destination, either explicitly designated by the user or implicitly supplied by the system (e.g. intermediate results in subexpressions or arguments for procedures). Some languages place the results of intermediate calculations in destinations beyond the control of the user. None the less, this standard defines the result of an operation in terms of that destination format as well as the values of the operands.

3.2.14 *Mode*

A mode is a variable which a program may set, sense, save and restore, to control the execution of subsequent arithmetic operations. The default mode is that mode which a