NORME INTERNATIONALE

INTERNATIONAL STANDARD

CEI
IEC
559

Deuxième édition
Second edition
1989-01

Arithmétique binaire en virgule flottante pour systèmes à microprocesseurs

Binary floating-point arithmetic for microprocessor systems

Numéro de référence
Reference number
CEI/IEC 559: 1989

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 (VEI), 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 VEI peuvent être obtenus sur demande.

Les termes et définitions figurant dans la présente publication ont été soit repris du VEI, 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 617 de la CEI: Symboles graphiques pour schémas.

Les symboles et signes contenus dans la présente publication ont été soit repris des Publications 27 ou 617 de la CEI, soit spécifiquement approuvés aux fins de cette publication.

Publications de la CEI établies par le même Comité d'Etudes

L'attention du lecteur est attirée sur le deuxième feuillet de la couverture, qui énumère les publications de la CEI préparées par le Comité d'Etudes qui a établi la présente 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 (IEV), 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 IEV will be supplied on request.

The terms and definitions contained in the present publication have either been taken from the IEV 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 617: Graphical symbols for diagrams.

The symbols and signs contained in the present publication have either been taken from IEC Publications 27 or 617, or have been specifically approved for the purpose of this publication.

IEC publications prepared by the same Technical Committee

The attention of readers is drawn to the back cover, which lists IEC publications issued by the Technical Committee which has prepared the present publication.

NORME INTERNATIONALE

INTERNATIONAL STANDARD

CEI
IEC
559

Deuxième édition
Second edition
1989-01

Arithmétique binaire en virgule flottante pour systèmes à microprocesseurs

Binary floating-point arithmetic for microprocessor systems

Bureau Central de la Commission Electrotechnique Internationale 3, rue de Varembé Genève, Suisse

Commission Electrotechnique Internationale
International Electrotechnical Commission
Международная Электротехническая Комиссия

CODE PRIX
PRICE CODE 18

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

SOMMAIRE

CONTENTS

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(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.

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.

2. Définitions

Exposant avec excédent

Somme de l'exposant et d'une constante (excédent ou biais) choisie de manière à rendre non négatif le domaine de l'exposant avec excédent.

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.

Nombre binaire 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 élevé à la puissance de son exposant. Dans la présente norme, une chaîne de bits n'est pas toujours distinguée du nombre qu'elle représente.

Nombre dénormalisé

Nombre en virgule flottante non nul dont l'exposant a une valeur réservée, d'habitude la valeur minimale du format et dont le bit significatif de la mantisse, explicite ou implicite, est nul.

Destination

Emplacement devant contenir le résultat d'une opération binaire ou unaire. La destination peut être soit désignée explicitement par l'utilisateur ou fournie de manière implicite par le système (pour les résultats intermédiaires dans les sous-expressions ou les arguments de procédures par exemple). Certains langages placent les résultats des calculs intermédiaires dans des emplacements non accessibles par l'utilisateur. Néanmoins, cette norme définit le résultat d'une opération en termes du format de cette destination aussi bien que des valeurs des opérandes.

Exposant

Elément d'un nombre binaire en virgule flottante qui représente normalement la puissance entière à laquelle deux est élevé pour déterminer la valeur du nombre représenté. Occasionnellement, l'exposant est appelé exposant signé ou exposant sans excédent.

Partie fractionnaire

Partie de la mantisse située à droite de sa virgule correspondante.

Mode

Variable qu'un utilisateur peut positionner, tester, sauvegarder et restaurer pour diriger l'exécution des opérations arithmétiques ultérieures. Le mode par défaut est le mode valable tant qu'une instruction contraire explicite n'est pas incluse dans le programme ou sa spécification.

Les modes suivants doivent être mis en place:

1) arrondi, pour commander la direction des erreurs d'arrondi, et dans certaines réalisations;

2) précision de l'arrondi, pour diminuer la précision des résultats. Le réalisateur peut fournir optionnellement les modes suivants:

3) déroutements désactivés/activés, pour gérer les conditions d'exception.

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 (non-nombre)

Non-nombre; entité symbolique codée selon le format virgule flottante. Il existe deux types de non-nombres (voir 6.2). Les non-nombres indicateurs indiquent une condition d'exception concernant une opération invalide (voir 7.1) lorsqu'ils apparaissent en tant qu'opérandes. Les non-nombres muets se propagent à travers presque toutes les opérations arithmétiques sans signaler d'exceptions.

Résultat

Chaîne de bits (représentant généralement un nombre) qui est livrée à la destination.

Mantisse

Elément d'un nombre binaire en virgule flottante constituée d'un bit significatif explicite ou implicite placé à gauche de la virgule et d'un champ fractionnaire à droite de la virgule.

Doit

Le mot "doit" recouvre les parties obligatoires de toute réalisation conforme.

Il convient - Il est recommandé - Il y a lieu

Ces termes recouvrent les parties qui sont fortement recommandées comme étant dans l'esprit de cette norme, bien que des contraintes architecturales ou autres, hors du domaine de cette norme, puissent à l'occasion rendre ces recommandations peu pratiques.

Indicateur d'état

Variable qui peut prendre deux états, actif (valeur 1) ou inactif (valeur 0). Un utilisateur peut désactiver un indicateur, le copier, ou le remettre dans un état antérieur. Lorsqu'il est actif, un indicateur peut contenir des informations supplémentaires dépendant du système et éventuellement inaccessibles à certains utilisateurs. Les opérations définies par cette norme peuvent avoir comme effet secondaire l'activation de certains des indicateurs suivants: résultat inexact, dépassement de capacité inférieur, dépassement de capacité supérieur, division par zéro et opération invalide.

Utilisateur

Toute personne, tout matériel ou logiciel non spécifié lui-même dans cette norme, ayant accès aux opérations de l'environnement de programmation spécifiées dans cette norme et qui les commande.

3. Formats

Cette norme définit quatre formats de virgule flottante en deux groupes, de base et étendu, chacun admettant deux largeurs, simple et double précision. Les niveaux de réalisation standard se distinguent par les combinaisons de formats supportés.

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 *Ensembles de valeurs*

Ce paragraphe concerne seulement les valeurs numériques représentables dans un format, et non leur codage qui fait l'objet des paragraphes suivants. Les seules valeurs représentables dans un format donné sont celles qui sont spécifiées selon les trois paramètres entiers suivants:

P = nombre de bits significatifs (précision)

E_{max} = valeur maximale de l'exposant, et

E_{min} = valeur minimale de l'exposant

Les paramètres de chaque format sont regroupés dans le tableau 1. Pour chaque format, les seules entités qui doivent être fournies sont:

Nombres de la forme $(-1)^s 2^E (b_0 b_1 b_2 \ldots b_{p-1})$

où:

s vaut 0 ou 1;
E est un entier compris entre E_{min} et E_{max} bornes incluses, et
chaque b_i vaut 0 ou 1.

Deux valeurs infinies, $+\infty$ et $-\infty$;
au moins un non-nombre indicateur, et
au moins un non-nombre muet.

Tableau 1 - Résumé des paramètres du format

Paramètre	Format			
	Simple	Simple Etendu	Double	Double Etendu
P	24	≥32	53	≥64
E_{max}	+127	≥+1 023	+1 023	≥+16 383
E_{min}	-126	≤-1 022	-1 022	≤-16 382
Excédent de l'exposant	+127	Non spécifié	+1 023	Non spécifié
Largeur de l'exposant (bits)	8	≥11	11	≥15
Largeur du format (bits)	32	≥43	64	≥79

La description précédente énumère certaines valeurs de manière redondante, par exemple:

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

Cependant, le codage de telles valeurs non nulles peut être redondant seulement pour les formats étendus (voir 3.3). Les valeurs non nulles de la forme $\pm 2^{E_{min}} (0 \cdot b_1 b_2 \ldots b_{p-1})$ sont appelées

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