

GUIDE : STANDARDS DE LA VIRGULE FLOTTANTE - IEEE STD 754-1985

FICHE N° 15310

PRÉSERVER
SAUVEGARDER
VALORISER

Période de fabrication : 1975-1999
Fabricant : fabricant non renseigné
Domaines : Informatique et Communication
Sous-domaines : Ordinateurs
Organisme : ACONIT
Ville : Grenoble (Isère)
Modèle :
Matériaux : Carton, Papier

Description

Titre traduit : IEEE Std 754-1985 - Floating-Point standard

Ce fascicule contient la norme internationale IEEE Std 754-1985 - Floating-Point standard qui définit, en 16 pages et 8 chapitres, les règles nécessaires pour effectuer sur ordinateur digital, et de manière reproductible, des calculs arithmétiques en « virgule flottante ».

Le terme virgule flottante caractérise le mode de codage des nombres dits « réels » qui seront manipulés dans les calculs arithmétiques. La représentation des nombres en virgule flottante consiste à séparer les chiffres significatifs et le facteur multiplicatif donnant l'ordre de grandeur. Ainsi le nombre 37,54 va être séparé en "3754" et "2" pour indiquer qu'il y a 2 chiffres avant la virgule décimale.

De façon plus mathématique cela revient à noter le nombre sous la forme $0,3754 \times 10$ puissance 2.

La norme IEEE-754 définit les standards à utiliser pour des « nombres réels » codées en « virgule flottante » (formats simple et double), avec des règles pour effectuer les arrondis (rounding) ou pour réaliser leur transformation en nombres entiers. Est aussi définie la manière dont ces nombres doivent être transmis entre programmes informatiques et manipulés (opérations) par les ordinateurs. Sont décrits aussi les cas d'exceptions et la gestion de cas éventuels d'erreur.

La norme IEEE-754-1985 est la première édition, datée de 1985, de la norme IEEE-754 à vocation internationale ; une seconde édition, modifiée et datant de 2008 a depuis été établie. Cette norme est publiée par l'IEEE (The Institut of Electrical and Electronics Engineers), un institut privé développant, à l'aide de contributions volontaires, des normes ayant vocation à être appliquées par toute l'industrie électronique. Ce fascicule simplifié est accompagné d'un fascicule détaillé (AC_25931)

Utilisation

Cette norme est destinée aux personnes qui programment des instructions de calcul au niveau machine (function's subroutines) des ordinateurs. En particulier, elle a été très tôt implémentée, partiellement ou totalement, dans des co-processeurs, c'est-à-dire des processeurs dédiés aux calculs et travaillant sous contrôle d'un processeur principal. Un bon exemple en est le co-processeur Intel_8087 (voir Aconit n° 20093 pour l'Intel i487 SX datant de 1989).

Avant 1985, des calculs en virgule flottante étaient déjà effectués avec des programmes informatiques réalisés selon des règles propres à chaque utilisateur ou à chaque constructeur d'ordinateur. Ce type de calcul doit être « reproductible », en particulier en cas de changement de machine ou de changement dans l'environnement informatique. Faute d'assurer cette reproductibilité, l'ordinateur risque de fournir des résultats différents, et d'autant plus erronés qu'il s'agira de gros calculs complexes et successifs.

Au tout début des années 1960, le programme spatial américain a eu besoin de réaliser de très gros calculs pour prévoir les trajectoires des fusées et des satellites. L'arrivée de l'ordinateur IBM-7090, suivi bientôt de son petit frère l'IBM-7044, visait à répondre à ce besoin, en permettant de gros calculs en virgule-flottante. Mais des erreurs ont été constatées dans des calculs de trajectoires. Un travail important de recherche a alors été lancé pour identifier les causes d'erreur et pour définir les modalités et conventions de codage des nombres, ainsi que les règles à appliquer pour les opérations. Ce sont les résultats de ces travaux de recherche qui ont été diffusés et rassemblés pour, en 1985, être pour la première fois standardisés dans cette norme IEEE-754. Cette norme souligne que l'influence d'Intel pour la définition technique de la virgule flottante a été importante.

An American National Standard

**IEEE Standard for Binary Floating-Point
Arithmetic**

Sponsor
**Standards Committee
of the
IEEE Computer Society**

Approved March 21, 1985
Reaffirmed December 6, 1990
IEEE Standards Board

Approved July 26, 1985
Reaffirmed May 21, 1991
American National Standards Institute

© Copyright 1985 by

The Institute of Electrical and Electronics Engineers, Inc
345 East 47th Street, New York, NY 10017, USA

No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher.

IEEE Standards documents are developed within the Technical Committees of the IEEE Societies and the Standards Coordinating Committees of the IEEE Standards Board. Members of the committees serve voluntarily and without compensation. They are not necessarily members of the Institute. The standards developed within IEEE represent a consensus of the broad expertise on the subject within the Institute as well as those activities outside of IEEE which have expressed an interest in participating in the development of the standard.

Use of an IEEE Standard is wholly voluntary. The existence of an IEEE Standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE Standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard. Every IEEE Standard is subjected to review at least once every five years for revision or reaffirmation. When a document is more than five years old, and has not been reaffirmed, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE Standard.

Comments for revision of IEEE Standards are welcome from any interested party, regardless of membership affiliation with IEEE. Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments.

Interpretations: Occasionally questions may arise regarding the meaning of portions of standards as they relate to specific applications. When the need for interpretations is brought to the attention of IEEE, the Institute will initiate action to prepare appropriate responses. Since IEEE Standards represent a consensus of all concerned interests, it is important to ensure that any interpretation has also received the concurrence of a balance of interests. For this reason IEEE and the members of its technical committees are not able to provide an instant response to interpretation requests except in those cases where the matter has previously received formal consideration.

Comments on standards and requests for interpretations should be addressed to:

Secretary, IEEE Standards Board
345 East 47th Street
New York, NY 10017
USA

Members of the Floating-Point Working Group of the Microprocessor Standards Subcommittee and those who participated by correspondence were as follows:

David Stevenson, Chair

Andrew Allison	Paul F. Flanagan	John C. Nash
William Ames	Gordon Force	Dan O'Dowd
Mike Arya	Lloyd Fosdick	Cash Olsen
Janis Baron	Robert Fraley	A. Padegs
Steve Baumel	Howard Fullmer	John F. Palmer
Dileep Bhandarkar	Daniel D. Gajski	Beresford Parlett
Joel Boney	David M. Gay	Dave Patterson
E.H. Bristol	C.W. Gear	Mary H. Payne
Werner Buchholz	Martin Graham	Tom Pittman
Jim Bunch	David Gustavson	Lew Randall
Ed Burdick	Guy K. Haas	Robert Reid
Gary R. Burke	Kenton Hanson	Christian Reinsch
Paul Clemente	Chuck Hastings	Frederic N. Ris
W.J. Cody	David Hough	Stan Schmidt
Jerome T. Coonen	John Edward Howe	Van Shahan
Jim Crapuchettes	Thomas E. Hull	Robert L. Smith
Itzhak Davidesko	Suren Irukulla	Roger Stafford
Wayne Davison	Richard E. James III	G.W. Stewart
R.H. Delp	Paul S. Jensen	Robert Stewart
James Demmel	W. Kahan	Harold S. Stone
Donn Denman	Howard Kaikow	W.D. Strecker
Alvin Despain	Richard Karpinski	Robert Swarz
Augustin A. Dubrulle	Virginia Klema	George Taylor
Tom Eggers	Les Kohn	James W. Thomas
Philip J. Faillace	Dan Kuyper	Dar-Sun Tsien
Richard Fateman	M. Dundee Maples	Greg Walker
David Feign	Roy Martin	John Steven Walther
Don Feinberg	William H. McAllister	Shlomo Waser
Smart Feldman	Colin McMaster	P.C. Waterman
Eugene Fisher	Dean Miller	Charles White
	Webb Miller	

When the IEEE Standards Board approved this standard on March 21, 1985, it had the following membership:

John E. May, Chair
John P. Riganati, Vice Chair
Sava I. Sherr, Secretary

James H. Beall	Jay Forster	Lawrence V. McCall
Fletcher J. Buckley	Daniel L. Goldberg	Donald T. Michael*
Rene Castenschiold	Kenneth D. Hendrix	Frank L. Rose
Edward Chelotti	Irvin N. Howell, Jr	Clifford O. Swanson
Edward J. Cohen	Jack Kinn	J. Richard Weger
Paul G. Cummings	Joseph L. Koepfinger*	W.B. Wilkens
Donald C. Fleckenstein	Irving Kolodny	Charles J. Wylie
	R.F. Lawrence	

***Member emeritus**

3.2 Basic Formats

Numbers in the single and double formats are composed of the following three fields:

- 1) 1-bit sign s
- 2) Biased exponent $e = E + \text{bias}$
- 3) Fraction $f = \cdot b_1 b_2 \dots b_{p-1}$

The range of the unbiased exponent E shall include every integer between two values $E_{\min}$ and $E_{\max}$, inclusive, and also two other reserved values $E_{\min}-1$ to encode ± 0 and denormalized numbers, and $E_{\max}+1$ to encode $\pm\infty$ and NaNs. The foregoing parameters are given in Table 1. Each nonzero numerical value has just one encoding. The fields are interpreted as follows:

3.2.1 Single

A 32-bit single format number X is divided as shown in Fig 1. The value v of X is inferred from its constituent fields thus

- 1) If $e = 255$ and $f \neq 0$, then v is NaN regardless of s
- 2) If $e = 255$ and $f = 0$, then $v = (-1)^s \infty$
- 3) If $0 < e < 255$, then $v = (-1)^s 2^{e-127} (1 \cdot f)$
- 4) If $e = 0$ and $f \neq 0$, then $v = (-1)^s 2^{-126} (0 \cdot f)$ (denormalized numbers)
- 5) If $e = 0$ and $f = 0$, then $v = (-1)^s 0$ (zero)

3.2.2 Double

A 64-bit double format number X is divided as shown in Fig 2. The value v of X is inferred from its constituent fields thus

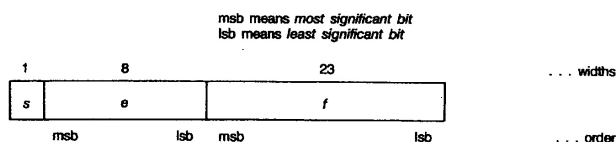


Figure 1— Single Format

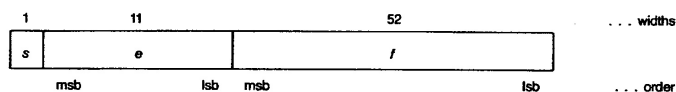


Figure 2— Double Format

- 1) If $e = 2047$ and $f \neq 0$, then v is NaN regardless of s
- 2) If $e = 2047$ and $f = 0$, then $v = (-1)^s \infty$
- 3) If $0 < e < 2047$, then $v = (-1)^s 2^{e-1023} (1 \cdot f)$
- 4) If $e = 0$ and $f \neq 0$, then $v = (-1)^s 2^{-1022} (0 \cdot f)$ (denormalized numbers)
- 5) If $e = 0$ and $f = 0$, then $v = (-1)^s 0$ (zero)



IEEE Standard for Floating-Point Arithmetic

IEEE Computer Society

Sponsored by the
Microprocessor Standards Committee

754TM

IEEE
3 Park Avenue
New York, NY 10016-5997, USA
29 August 2008

IEEE Std 754TM-2008
(Revision of
IEEE Std 754-1985)

Authorized licensed use limited to: University of Maryland Baltimore Cty. Downloaded on September 03, 2015 at 19:44:10 UTC from IEEE Xplore. Restrictions apply.

Pour nous citer :

Base de la Mission nationale de sauvegarde et de valorisation du patrimoine scientifique et technique contemporain, PATSTEC, Guide : Standards de la virgule flottante - IEEE Std 754-1985 (fabricant non renseigné),
<https://www.patstec.fr/ressources/objets/detail?id=29585>, consulté le 2026-07-25