

# IEC 559:1989-01 (E/F)

## Binary floating-point arithmetic for microprocessor systems

## Arithmétique binaire en virgule flottante pour systèmes à microprocesseurs

---

### SOMMAIRE

	Pages
PREAMBULE .....	4
PREFACE .....	4
<b>Articles</b>	
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
<b>Clause</b>	
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