

## APPENDIX O



### FLOATING POINT FORMATS

<u>Paragraph</u>	<u>Title</u>	<u>Page</u>
1.0	Introduction.....	O-1
2.0	IEEE 32 Bit Single Precision Floating Point .....	O-1
3.0	IEEE 64 Bit Double Precision Floating Point .....	O-2
4.0	MIL STD 1750A 32 Bit Single Precision Floating Point.....	O-2
5.0	MIL STD 1750A 48 Bit Double Precision Floating Point .....	O-2
6.0	DEC 32 Bit Single Precision Floating Point.....	O-3
7.0	DEC 64 Bit Double Precision Floating Point .....	O-3
8.0	IBM 32 Bit Single Precision Floating Point .....	O-3
9.0	IBM 64 Bit Double Precision Floating Point.....	O-4
10.0	TI (Texas Instruments) 32 Bit Single Precision Floating Point.....	O-4
11.0	TI (Texas Instruments) 40 Bit Extended Precision Floating Point.....	O-4

### LIST OF TABLES

Table O-1.	Floating Point Formats.....	O-1
------------	-----------------------------	-----

This page intentionally left blank.

## APPENDIX O

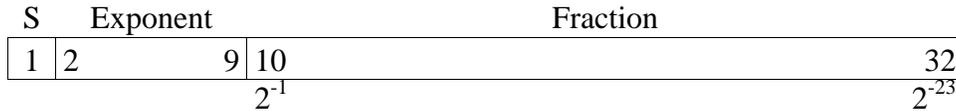
### FLOATING POINT FORMATS

#### 1.0 Introduction

The following table provides a summary of the floating point formats listed in Table O-1. Details of each format are provided in the paragraphs following the table.

TABLE O-1. FLOATING POINT FORMATS							
Type	Size	Radix	Sign	Exponent	Fraction	Bias	Formula
IEEE_32	32	2	1	8	23	127	$(-1^S)(1.F)(2^{(E-127)})$
IEEE_64	64	2	1	11	52	1023	$(-1^S)(1.F)(2^{(E-1023)})$
1750A_32	32	2	0	8	24	0	$(0.F)(2^E)$
1750A_48	48	2	0	8	40	0	$(0.F)(2^E)$
DEC_32	32	2	1	8	23	128	$(-1^S)(0.1F)(2^{(E-128)})$
DEC_64	64	2	1	8	55	128	$(-1^S)(0.1F)(2^{(E-128)})$
IBM_32	32	16	1	7	24	64	$(-1^S)(0.F)(16^{(E-64)})$
IBM_64	64	16	1	7	56	64	$(-1^S)(0.F)(16^{(E-64)})$
TI_32	32	2	1	8	24	0	$((-2)^S + (0.F))(2^E)$
TI_40	40	2	1	8	32	0	$((-2)^S + (0.F))(2^E)$

#### 2.0 IEEE 32 Bit Single Precision Floating Point



$$\text{Value} = (-1^S)(1.F)(2^{(E-127)})$$

S = sign: 0 = Positive, 1 = Negative  
 Exponent = power of 2 with bias of 127  
 Fraction = F portion of 23 bit fraction 1.F  
 0: E = 0, F = 0

### 3.0 IEEE 64 Bit Double Precision Floating Point

S	Exponent		Fraction	
1	2	12	13	64
	$2^{-1}$		$2^{-1}$	$2^{-52}$

$$\text{Value} = (-1^S)(1.F)(2^{(E-1023)})$$

S = sign: 0 = Positive, 1 = Negative  
 Exponent = power of 2 with bias of 1023  
 Fraction = F portion of 52 bit fraction 1.F  
 0: E = 0, F = 0

### 4.0 MIL STD 1750A 32 Bit Single Precision Floating Point

S	Fraction		Exponent	
1	2	24	25	32
	$2^{-1}$		$2^{-23}$	

$$\text{Value} = (0.F)(2^E)$$

Exponent = 2's complement power of 2  
 Fraction = Normalized, 2's complement F portion of 24 bit fraction 0.F  
 (Bit 2 MUST be set for positive, clear for negative)  
 0: F = 0

### 5.0 MIL STD 1750A 48 Bit Double Precision Floating Point

S	Fraction (MSW)		Exponent		Fraction	
1	2	24	25	32	33	48
	$2^{-1}$		$2^{-23}$		$2^{-24}$	$2^{-31}$

$$\text{Value} = (0.F)(2^E)$$

Exponent = 2's complement power of 2  
 Fraction = Normalized, 2's complement F portion of 40 bit fraction 0.F  
 (Bit 2 MUST be set for positive, clear for negative)  
 0: F = 0

**6.0 DEC 32 Bit Single Precision Floating Point**

S	Exponent		Fraction	
1	2	9	10	32
			$2^{-2}$	$2^{-24}$

$$\text{Value} = (-1^S)(0.1F)(2^{(E-128)})$$

S = sign: 0 = Positive, 1 = Negative  
 Exponent = power of 2 with bias of 128  
 Fraction = F portion of 23 bit fraction 0.1F  
 0: S = 0 & F = 0 & E = 0

**7.0 DEC 64 Bit Double Precision Floating Point**

S	Exponent		Fraction	
1	2	9	10	64
			$2^{-2}$	$2^{-56}$

$$\text{Value} = (-1^S)(0.1F)(2^{(E-128)})$$

S = sign: 0 = Positive, 1 = Negative  
 Exponent = power of 2 with bias of 128  
 Fraction = F portion of 55 bit fraction 0.1F  
 0: S = 0 & F = 0 & E = 0

**8.0 IBM 32 Bit Single Precision Floating Point**

S	Exponent		Fraction	
1	2	8	9	32
			$2^{-1}$	$2^{-24}$

$$\text{Value} = (-1^S)(0.F)(16^{(E-64)})$$

S = sign: 0 = Positive, 1 = Negative  
 Exponent = power of 16 with bias of 64  
 Fraction = Normalized F portion of 24 bit fraction 0.F  
 (Bits 9-12 cannot be all zero)  
 0: F = 0

**9.0 IBM 64 Bit Double Precision Floating Point**

S	Exponent	Fraction	
1	2	8	9
		$2^{-1}$	
			64
			$2^{-56}$

$$\text{Value} = (-1^S)(0.F)(16^{(E-64)})$$

S = sign: 0 = Positive, 1 = Negative  
 Exponent = power of 16 with bias of 64  
 Fraction = Normalized F portion of 56 bit fraction 0.F  
 (Bits 9-12 cannot be all zero)  
 0: F = 0

**10.0 TI (Texas Instruments) 32 Bit Single Precision Floating Point**

Exponent	S	Fraction	
1	8	9	10
		$2^{-1}$	
			32
			$2^{-23}$

$$\text{Value} = ((-2)^S + (0.F))(2^E)$$

Exponent = 2's complement power of 2  
 Fraction = 2's complement F portion of 24 bit fraction 1.F  
 0: E = -128

**11.0 TI (Texas Instruments) 40 Bit Extended Precision Floating Point**

Exponent	S	Fraction	
1	8	9	10
		$2^{-1}$	
			40
			$2^{-31}$

$$\text{Value} = ((-2)^S + (0.F))(2^E)$$

Exponent = 2's complement power of 2  
 Fraction = 2's complement F portion of 32 bit fraction 1.F  
 0: E = -128

\*\*\*\* END OF APPENDIX O \*\*\*\*