APPENDIX 9-D

Floating Point Formats

D.1. Introduction

Table D-1 provides a summary of floating point formats. Details of each format are shown on the pages following the table.

Table D-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	$(-1^{S})(0.1F)(2^{(E-128)})$
DEC_64G	64	2	1	11	52	1024	$(-1^{S})(0.1F)(2^{(E-1024)})$
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	$((-2)^{S} + (0.F))(2^{E})$

D.2. IEEE 754 32-Bit Single Precision Floating Point

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

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

where 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

D.3. IEEE 754 64-Bit Double Precision Floating Point

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

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

where 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

D.4. MIL-STD-1750A 32-Bit Single Precision Floating Point

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

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

where Exponent = 2's complement power of 2
S = sign: 0 = Positive, 1 = Negative
S + 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

D.5. MIL-STD-1750A 48-Bit Double Precision Floating Point

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

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

where Exponent = 2's complement power of 2
S = sign: 0 = Positive, 1 = Negative
S + 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

D.6. DEC 32-Bit Single Precision Floating Point

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

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

where 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

D.7. DEC 64-Bit Double Precision Floating Point

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

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

where 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

D.8. DEC 64-Bit "G" Double Precision Floating Point

S	Exponent		Fraction	
1	2	12	13	64
			2^{-2}	2^{-53}

Value = $(-1^S)(0.1F)(2^{(E-1024)})$

where S = sign: 0 = Positive, 1 = Negative
Exponent = power of 2 with bias of 1024
Fraction = F portion of 52-bit fraction 0.1F
0: S = 0 & F = 0 & E = 0

D.9. IBM 32-Bit Single Precision Floating Point

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

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

where 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

D.10. IBM 64-Bit Double Precision Floating Point

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

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

where 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

D.11. TI (Texas Instruments) 32-Bit Single Precision Floating Point

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

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

where Exponent = 2’s complement power of 2
S = sign: 0 = Positive, 1 = Negative
Fraction = 2’s complement F portion of 24-bit fraction 1.F
0: E = −128

D.12. TI (Texas Instruments) 40-Bit Extended Precision Floating Point

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

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

where Exponent = 2’s complement power of 2
S = sign: 0 = Positive, 1 = Negative
Fraction = 2’s complement F portion of 32-bit fraction 1.F
0: E = −128