Engineering Notebook

miscellaneous problems and solutions

compiled by C. Bond Vol.1 no.15

Machine Constants for IEEE Math

The following numbers are the fundamental constants conforming to the IEEE floating point math specification. In particular, these values pertain to floating point processors for the IBM pc. See IMACH for definitions of B, T, EMIN and EMAX. The names RMACH, DMACH and IMACH are from the FORTRAN portable code conventions described in the reference.

Name	Value	Description
RMACH(1)	B^ (EMIN-1)	the smallest possible magnitude
RMACH(2)	B^ EMAX×(1-B^ (-T))	the largest magnitude
RMACH(3)	B^ (-T)	the smallest relative spacing
RMACH(4)	B^ (1-T)	the largest relative spacing
RMACH(5)	LOG_{10}(B)	log base 10 of the radix

In floating point notation (for 4-byte reals) :

Name	Scientific Notation	Radix Notation
RMACH(1)	$1.17549435e-38$	2^ -126
RMACH(2)	$3.40282347e+38$	2^ 128×(1-2^ 24)
RMACH(3)	$5.95604645e-8$	2^ -24
RMACH(4)	$1.19209290e-7$	2^ -23
RMACH(5)	0.30102999566	$\mathrm{LOG}_{10}2$

In floating point notation (for 8-byte reals) :

Name	Scientific Notation	Radix Notation
DMACH(1)	$2.2250738585072014e-308$	2^ -1022
DMACH(2)	$1.7976931348623157e+308$	2^ 1024×(1-2^ -53)
DMACH(3)	$1.1102230246251565e-16$	2^ -53
DMACH(4)	$2.2204460492503131e-16$	2^ -52
DMACH(5)	0.30102999566398119 5	$LOG_{10}2$

Assume integers are represented in the 8-digit, base-A form

$$\text{sign}(X(S-1)\times A^{S-1}+\cdots+X(1)\times A+X(0))$$

where $0 \le X(I) < A$ for $I = 0, \ldots, S-1$,

Then for 2-byte integers,

Name	Symbol	Value	Description
IMACH(7)	A	2	the base
IMACH(8)	S	15	the number of base-A digits
IMACH(9)	AŜ-1	32767	the largest magnitude

For 4-byte integers,

Name	Symbol	Value	Description
IMACH(7)	A	2	the base
IMACH(8)	S	31	the number of base-A digits
IMACH(9)	AŜ-1	2147483647	the largest magnitude

Assume floating-point numbers are represented in the T-digit base-B form

$$\text{sign}(B^E)\times((X(1)/B)+\cdots+(X(T)/B^T))$$

where

$$0 \le X(I) < B \text{ for } I = 1, \ldots, T \tag{1}$$

$$0 < X(I), \text{ and } EMIN \le E \le EMAX. \tag{2}$$

The base for all floating point numbers is:

Name	Symbol	Value	Description
IMACH(10)	B	2	the base

For single precision floats,

Name	Symbol	Value	Description
IMACH(11)	T	24	the number of base-B digits
IMACH(12)	EMIN	-125	the smallest exponent E
IMACH(13)	EMIN	128	the largest exponent E

For double precision floats,

Name	Symbol	Value	Description
IMACH(14)	T	53	the number of base-B digits
IMACH(15)	EMIN	-1021	the smallest exponent E
IMACH(16)	EMIN	1024	the largest exponent E