CSARCH Lecture Series: Double precision floating-point format for decimal (Decimal 64)

Sensei RL Uy

College of Computer Studies

De La Salle University

Manila, Philippines

Copyright Notice

Overview

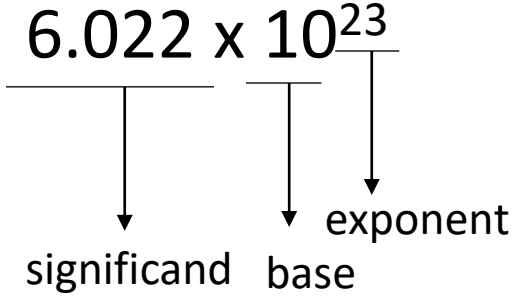
- This sub-module introduces the IEEE-754 decimal-64 floating-point format
- The objective is as follows:
  - ✓Describe the process of representing decimal-64 floating-point data using IEEE-754 standard

Floating Point

- Scientists and engineers use scientific notation where a number is expressed as

$$+/- S \times 10^{\pm E}$$

- Where S is the significand (also known as mantissa), E is the exponent and 10 is the base.



Floating Point

- Floating point standard for floating-point numbers in computer is the IEEE-754 (Institute of Electrical and Electronics Engineers Standard 754). Originally 1985, revised 2008, current version 2019
- Decimal floating point is introduced in 2008.
- The representation is to be used in applications that need to emulate decimal rounding exactly (i.e., financial and tax computations)

IEEE-754 decimal floating-point

- Decimal32 precision
- Decimal64 precision
- Decimal128 precision

Length of field

Format	Decimal32	Decimal64	Decimal128
Format length	32	64	128
Sign bit	1	1	1
Combination bit	5	5	5
Exponent continuation bit	6	8	12
coefficient continuation bit	20	50	110
Total coefficient in digits	7	16	34
E_{max} (denormalized/normalized)	96/90	384/369	6144/6111
E_{min} (denornamlized/normalized)	-95/-101	-383/-398	-6143/-6176
Bias	101	398	6176
E_{limit}	191	767	12287

IEEE-754 decimal-64 floating-point format

Sign	Combination field	Exponent continuation	Coefficient continuation
1	5	8	50

normalized to this format before representation:

ddddddddddddddddd x 10^{e}

- IEEE-754 decimal64 floating-point format is 64-bit in width
- The 32-bit is partition as 1 bit for sign bit; 5 bits for combination field, 8 bits for exponent continuation and 50 bits for coefficient continuation
  - Significand in decimal
  - Base 10
  - sign bit: 0→ positive; 1→ negative
  - e’ = e+398
  - significand normalized to 16 whole decimal digits

Combination field

Combination Field	Type	Exp MSbs	coefficient MSD
a b c d e	Finite	a b	0 c d e
1 1 c d e	Finite	c d	1 0 0 e
1 1 1 1 0	Infinity	- -	- - - -
1 1 1 1 1	NaN	- -	- - - -

- 5-bit combination field is composed of:
  - two most significant bits of the exponent representation (valid bits: 00, 01 and 10 only)
  - 1 or 3 bits of the most significant digit of the significand

Exponent continuation field

- Exponent representation is e+398
  - two most significant bits of the exponent representation (valid bits: 00, 01 and 10 only) in the combination field
  - The rest of the 8 bits in the exponent continuation field
  - Largest exponent value that can be represented is 384
  - smallest exponent value that can be represented is -383

Coefficient continuation field

- 16 whole decimal digits
- most significant digit is stored in the combination field
- remaining 15 digits are represented as densely-packed BCD and stored in the coefficient continuation field

Example

- $7531123456574426 \times 10^{20}$

Combination Field	Type	Exp MSBs	coefficient MSD
a b c d e	Finite	a b	0 c d e
1 1 c d e	Finite	c d	1 0 0 e
1 1 1 1 0	Infinity	- -	- - - -
1 1 1 1 1	NaN	- -	- - - -

Significand in decimal?	yes
Base-10?	yes
Normalized?	Yes, 16 whole digits
	MSD = 7 (0111)
Sign bit	0 (+)
e' = e+398	20+398 = 418 (01 10100010)

Sign	Combination field	Exponent continuation	Coefficient continuation
0	01 111	1010 0010	1010110001 0010100011 1001010110 1011110100 1000100110

Example

- $-8765432345678100 \times 10^{-20}$

Combination Field	Type	Exp MSBs	coefficient MSD
a b c d e	Finite	a b	0 c d e
1 1 c d e	Finite	c d	1 0 0 e
1 1 1 1 0	Infinity	- -	- - - -
1 1 1 1 1	NaN	- -	- - - -

Significand in decimal?	yes
Base-10?	yes
Normalized?	Yes, 16 whole digits
	MSD = 8 (1000)
Sign bit	1 (-)
e’ = e+101	-20+398 = 378 (01 01111010)

Sign	Combination field	Exponent continuation	Coefficient continuation
1	11 010	01111010	1111100101 100011010 0111000101 1101111000 0010000000

Try

- -1.234567×10^{15}

Combination Field	Type	Exp MSBs	coefficient MSD
a b c d e	Finite	a b	0 c d e
1 1 c d e	Finite	c d	1 0 0 e
1 1 1 1 0	Infinity	- -	- - - -
1 1 1 1 1	NaN	- -	- - - -

Significand in decimal?	
Base-10?	
Normalized?	
Sign bit	
e’ = e+101	

Sign	Combination field	Exponent continuation	Coefficient continuation

Try

- -1.234567×10^{15}

Combination Field	Type	Exp MSbs	coefficient MSD
a b c d e	Finite	a b	0 c d e
1 1 c d e	Finite	c d	1 0 0 e
1 1 1 1 0	Infinity	- -	- - - -
1 1 1 1 1	NaN	- -	- - - -

Significand in decimal?	yes
Base-10?	yes
Normalized?	No, -0000000001234567 x 10^9
	MSD = 0 (0000)
Sign bit	1 (-)
e' = e+101	9+398 = 407 (01 10010111)

Sign	Combination field	Exponent continuation	Coefficient continuation
1	01 000	10010111	0000000000 0000000000 0000000001 0100110100 1011100111

To recall …

- What have we learned:
  - ✓Describe the process of representing decimal-64 floating-point data using IEEE-754 standard