Computer Organization

Floating Point Part II

Prof. Roger Luis Uy
De La Salle University
College of Computer Studies

Decimal Floating Point Representation

- Decimal representation matches the definition of decimal numbers used in almost all databases, programming languages and applications
- Thus, IEEE-754-2008 is introduced in 2008 to support decimal FP representation

Decimal Floating Point Representation

- There are 3 representation: Decimal32 (32-bit), Decimal64 (64-bit) & Decimal128 (128-bit)

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
Mantissa continuation bit	20	50	110
Total mantissa in digits	7	16	34
E_{max}	96	384	6144
E_{min}	-95	-383	-6143
Bias	101	398	6176
E_{limit}	191	767	12287

Length of Field

Format	Decimal32	Decimal64	Decimal128
Largest value	$9.99..x10^{\text{Emax}}$	$9.99..x10^{\text{Emax}}$	$9.99..x10^{\text{Emax}}$
Smallest value	$1.00..x10^{\text{Emin}}$	$1.00..x10^{\text{Emin}}$	$1.00..x10^{\text{Emin}}$
Smallest non-zero	$1.00..x10\ E^{-\text{bias}}$	$1.00..x10\ E^{-\text{bias}}$	$1.00..x10\ E^{-\text{bias}}$

Example: for Decimal 32

Largest value: $9.999999x10^{96} = 9999999x10^{90}$

Smallest value: $1.000000x10^{-95} = 1000000x10^{-101}$

Smallest subnormal value: $0.000001\ x10^{-95} = 1*10^{-101}$

Example: for Decimal 64

Largest value: $9.999999999999999x10^{384} = 9999999999999999x10^{369}$

Smallest value: $1.000000000000000x10^{-383} = 1000000000000000x10^{-398}$

Smallest subnormal value: $0.000000000000001\ x10^{-383} = 1*10^{-398}$

1-bit sign bit

- 0 → positive
- 1 → negative

5-bit combination field

- Two most significant bits of the exponent (value should be 00,01 and 10 only) why?
- (1 or 3 bits) Most significant digit

Combination Field	Type	Exp MSBs	Mantissa 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	- -	- - - -

Exponent field

- Exponent to be represented is "biased" (e.g. 101 for Decimal32, 398 for Decimal64, 6176 for Decimal128)
- Two most significant bits of the exponent is place in the combination field
- The rest is in the exponent continuation length (*ecbits*)

Coefficient/Mantissa field

- Coefficient/Mantissa is represented as densely packed decimal encoding
- Densely packed decimal encoding is compressed 3 BCD digits to 10-bit
- Most significant digit is assigned to the combination field
- The rest is assigned to the coefficient/mantissa field

Redundant encoding

- 7.50

$= 750 \times 10^{-2}$

$= 7500 \times 10^{-3}$

$= 7500 \times 10^{-4}$

$= 75000 \times 10^{-5}$

$= 750000 \times 10^{-6}$

All are the same representation of 7.5

Example

- -7.50 represent in decimal 64 format
- = -750 x10^{-2}
- Exponent representation = -2+398 = 396 = 01 1000 1100 (in binary)
- Mantissa = 0000000000000750
- Sign bit = 1
- Combination field: 01 000
- Exponent continuation = 1000 1100
- Mantissa continuation = 0000000000 0000000000 0000000000 0000000000 111 101 0 000

Example

- 7.25 x10^5 represent in decimal 32 format
- = 725 x10^3
- Exponent representation = +3+101 = 104 = 01 101 000 (in binary)
- Mantissa = 0 000725
- Sign bit = 0
- Combination field: 01 000
- Exponent continuation = 101 000
- Mantissa continuation = 0000000000 111 101 0101