REAL NUMBERS in BINARY NOTATION

by

Prof. Dr. M. Akram Tahir

1

SEM NOTATION

- Real numbers are stored in SEM notation.
- S stands for Sign, E for Exponent and M for Mantissa.
- There would be three parts of the number in storage: SIGN, EXPONENT and MANTISSA.

2

SIGN

- The highest bit will denote the sign of the real number. A ONE in this bit indicates Negative Number and a ZERO denotes a Non-negative Number.

3

EXPONENT

- Next to sign bit would be a set of n bits containing EXPONENT.
- Exponent is abbreviated by EXP.
- EXP is stored in EXCESS Notation.
- EXP is used as power of 2.

4

MANTISSA

- Mantissa is the part of the number which (after adjusting) will be multiplied by 2^{EXP} to get the number.
- Mantissa is stored in FRACTIONAL notation.

5

FRACTIONAL NOTATION

- Note the decimal number:

  534.162
- The number is obtained as:
- $5x10^{2}+3x10^{1}+4x10^{0}+1x10^{-1}+6x10^{-2}+2x10^{-3}$
- That is left of radix point local values are decreased by multiple of 1/10

6

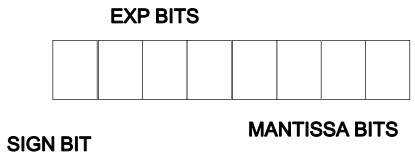
FRACTIONAL NOTATION

- If Mantissa is stored in N bits, the local value of Nth bit would be 1.
- The local value of 1st bit would be $2^{-(N-1)}$.
- The Local value of Ith bit (from low side) would be $2^{-(N-I)}$.

7

REAL in 1 BYTE

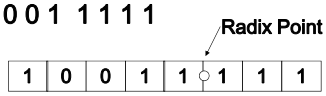
❖ This example is of a hypothetical real number in 1+3+4 SEM storage.



8

BIT PATTERN

❖ 1 0 0 1 1 1 1 1



❖Sign is 1 (-ve)

❖EXP is 001 in E4 notation (-3)

❖Mantissa is 1111 in Fractional notation = 1.875

9

FRACTIONAL NOTATION

- 0 001 1111
- Mantissa : 1111

$= 2^0+2^{-1}+2^{-2}+2^{-3}$

$= 1+1/2+1/4+1/8$

$= 1.875$

10

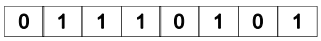
RULE (i)

1. If EXP is the smallest number, it will be adjusted by adding +1 and the following formula will be used for obtaining the real value:
2. Real Number = $S\ 2^{\text{Adjusted EXP}} \times$ MANTISSA (Formula 1)

11

RULE (ii)

2. If EXP is the largest number it will denote the OVERFLOW.



1 1 1 = +3 in E4

12

RULE (iii)

3) For all other valid EXPs the following formula is used for obtaining the real value:

Real Number = S 2^{EXP} x (2 + MANTISSA)

(Formula 2)

i.e., 2 is added to the Mantissa.

13

EXAMPLE 1

- In SEM 1+3+4 determine the real number from 0101 0111
- Sign is (0) +ve
- EXP is 101 = 1 in E4 Notation
- Mantissa is 0111 = 0+1/2+1/4+1/8=7/8
- Hence the number is $+2^{1}(2+7/8) = 5.75$

14

EXAMPLE 2

- In SEM 1+3+4 determine the real number from 0000 1111
- Sign is +ve
- EXP is 000 = - 4 in E4 Notation
- Mantissa is 1111 = 1+1/2+1/4+1/8=15/8
- Hence number is $+2^{-3}$ (15/8) = 15/64

= 0.234375

15

EXAMPLE 3

- In SEM 1+3+4 determine the real number from 0001 1111
- Sign is +ve
- EXP is 001 = - 3 in E4 Notation
- Mantissa is 1111 = 1+1/2+1/4+1/8=15/8
- Hence number is $+2^{-3}(2+15/8) = 31/64$

= 0.484375

16

EXAMPLE 4

- In SEM 1+3+4 determine the real number from 0111 0011
- Sign is +ve
- EXP is 111 = + 3 in E4 Notation

Which indicates the OVERFLOW

No Real Number but Overflow

17

EXAMPLE 5

- In SEM 1+3+4 determine the real number from 1001 1111
- Sign is -ve
- EXP is 001 = - 3 in E4 Notation
- Mantissa is 1111 = 1+1/2+1/4+1/8=15/8
- Hence number is $- 2^{-3}(2+15/8) = -31/64$

= - 0.484375

18

EXAMPLE 6

- In SEM 1+3+4 determine the real number from 1101 1000
- Sign is -ve
- EXP is 101 = 1 in E4 Notation
- Mantissa is 1000 = 1
- Hence the number is - $2^1(2+1)$ = -6.0

19

EXAMPLE 7

- In SEM 1+3+4 determine the real number from 0101 1000
- Sign is +ve
- EXP is 101 = 1 in E4 Notation
- Mantissa is 1000 = 1
- Hence the number is $2^1(2+1)$ = +6.0

20

EXAMPLE 8

- In SEM 1+3+4 determine the real number from 0100 1111.
- Sign is +ve
- EXP is 100 = 0 in E4 Notation
- Mantissa is 1111 = 1+1/2+1/4+1/8=15/8
- Hence the number is $2^0(2+15/8)$ = +3.875

21

EXAMPLE 9

- In SEM 1+3+4 determine the real number from 0101 0000
- Sign is +ve
- EXP is 101 = 1 in E4 Notation
- Mantissa is 0000 = 0
- Hence the number is $2^1(2+0)$ = +4.0

22

REAL DATA TYPES

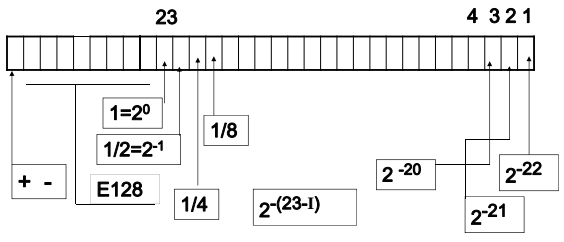
- SINGLE (Single Precision Numbers)
- 4 Byte storage in 1+8+23 SEM Notation
- The precision is 7 digits

- DOUBLE (Double Precision Numbers)
- 8 Byte storage in 1+11+52 SEM Notation
- The precision is 15 digits

23

SINGLE PRECISION

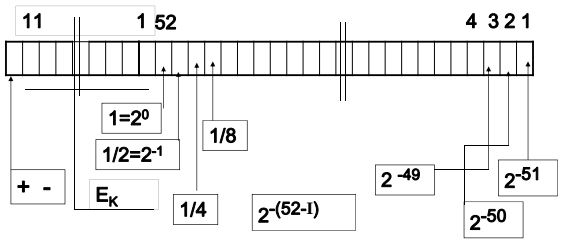
- SEM 1+8+23 in 4 BYTE WORD



24

DOUBLE PRECISION

- SEM 1+11+52 in 8 BYTE WORD



25

EXAMPLE 1 SINGLE

0000 0000 0100 0000 0000 0000 0000 0000

SIGN= +ve

EXP = -128

Mantissa = 1

Number = 2^{-127} (1) = 5.877472×10^{-39}

26

EXAMPLE 2 SINGLE

0000 0000 0000 0000 0000 0000 0000 0001

SIGN= +ve

EXP = -128

Mantissa = 2^{-22}

Number = 2^{-127} (2^{-22}) = 1.401298×10^{-45}

The Smallest +ve Number

27

EXAMPLE 3 SINGLE

1000 0000 0000 0000 0000 0000 0000 0001

SIGN= -ve

EXP = -128

Mantissa = 2^{-22}

Number = -2^{-127} (2^{-22}) = $-1.401298 \times 10^{-45}$

The Largest -ve Number

28

EXAMPLE 4 SINGLE

0111 1111 0111 1111 1111 1111 1111 1111

SIGN= +ve

EXP = 126 (In E128)

Mantissa = $1+ 2^{-1} + 2^{-2} + .. + 2^{-20} + 2^{-21} + 2^{-22}$

Number = 2^{126} (2+1.999999)

= 3.402823×10^{38} (The largest +ve Number)

29

EXAMPLE 5 SINGLE

1111 1111 0111 1111 1111 1111 1111 1111

SIGN= -ve

EXP = 126 (In E128)

Mantissa = $1+ 2^{-1} + 2^{-2} + .. + 2^{-20} + 2^{-21} + 2^{-22}$

Number = -2^{126} (2+1.999999)

= -3.402823×10^{38} (The smallest -ve Number)

30

EXAMPLE 6

SINGLE

1001 0010 0000 0100 0000 0000 0000 0001

SIGN= -ve

EXP = -92 (In E128)

Mantissa = $2^{-4} + 2^{-22}$

Number = $-2^{-92} (2 + 2^{-4} + 2^{-22})$

= $- 4.165186 \times 10^{-28}$

31

EXAMPLE 7

DOUBLE

0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0001

SIGN= +ve

EXP = -1024

Mantissa = 2^{-51}

Number = $2^{-1023} (2^{-51})$ = Smallest +ve Number

32

EXAMPLE 8

DOUBLE

0111 1111 1110 1111 1111 1111 1111 1111 1111 1111 1111 1111 1111 1111 1111 1111

SIGN= +ve

EXP = 1022

Mantissa = $1+2^{-1} + + + 2^{-51}$

Number = 2^{1022} (2+1.999 9) = Largest Number

33

EXAMPLE 9

DOUBLE

1000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0001

SIGN= -ve

EXP = -1024

Mantissa = 2^{-51}

Number = $-2^{-1023} (2^{-51})$ = largest -ve Number

34

EXAMPLE 10

DOUBLE

1111 1111 1110 1111 1111 1111 1111 1111 1111 1111 1111 1111 1111 1111 1111 1111

SIGN= -ve

EXP = 1022

Mantissa = $1+2^{-1} + + + 2^{-51}$

Number = -2^{1022} (2+1.999 9) = smallest -ve Number

35

36