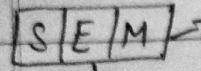
* IEEE 754 Notations Lecture 13 Part 1

- Single Precision
  32 bit (4 byte)
- double precision
  64 bit (8 byte)

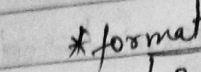
* format



0 +ve
1 -ve

8 bit
ex-127

* format



0 +ve
1 -ve

11 bit
ex-1023

* IEEE Normalized

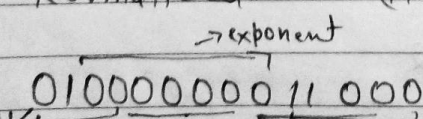
1. xxx.....

do not store
(implicit)

ex: $(3.5)_{10}$ → Convert it into ieee single precision normalized number

Step 1 => Binary = $(11.1)_2$

Step 2 => Normalized = $(11.1)_2$ => $1.11 * 2^1$



40600000H Ans

* $(0.5)_{10} \Rightarrow (0.1)_2 \Rightarrow 1.0 * 2^{-1}$

00111111000...

3F000000H Ans

* $(6.25)_{10} \Rightarrow (110.01)_2 \Rightarrow 1.1001 * 2^2$

01000000110010...

40C8000H Ans

* 3FC0000H → IEEE Single Precision, decimal?

0 01111111 1000000...

S E M

$+ 1.1 * 2^{127-127} \Rightarrow 1.1 * 2^0 \Rightarrow (1.1)_2$

$(1.5)_{10}$ Ans

* 40C80000H → IEEE Single Precision, decimal?

0 10000001 1001000...

E M

$+ 1.1001 * 2^{129-127} \Rightarrow 1.1001 * 2^2 \Rightarrow (110.01)_2$

$(6.25)_{10}$ Ans

* Denormals ⇒ exponent all 0's

if exponent ~~is~~ all !=0 means it is normalized

"1" is not implicit

* Ex: S e m
0 00000000 0000....

- all exponent 0 means it is denormalized and "1" is not implicit.

0 00000000 000...

$+ \quad 0 * 2^{0-127} \Rightarrow +0$

- 1 00000000 000...

$- \quad 0 * 2^{0-127} = -0$ Ans

- 00000000H ⇒ +0
80000000H ⇒ -0

* These two are special number

S E M

- 0 11111111 000 ---

$+\infty$

7F800000H = $+\infty$ => +ve infinity

S E M

- 1 11111111 000 ---

FF800000H => $-\infty$ => -ve infinity

both are special number

- 01111111 $\neq 0$ --- Not a number

11111111 $\neq 0$ --- Not a number

* S.P.N	S	E	M
+0	0	all 0's	all 0's
-0	1	all 0's	all 0's
$+\infty$	0	all 1's	all 0's
$-\infty$	1	all 1's	all 0's
NAN	0/1	all 1's	$\neq 0$
Denormels	0/1	all 0's	$\neq 0$

Largest * Normalized number

Part 2

- Largest +ve

S E M

0 | 1111 1111 | 0 1111 ... 1

7 F7FFFFF H Ans

$1.1111\ldots * 2^{254-127} \Rightarrow 2-2^{-23} * 2^{127}$

- Smallest +Ve

S E M

0 00000001 0000 ... ...

0 0800000 H Ans

$1.0 * 2^{1-127} \Rightarrow 1.0 * 2^{-126} \Rightarrow 2^{-126}$ Ans

- Smallest -Ve

$-(2-2^{-23}) * 2^{127}$

FF7FFFFF H Ans

- Largest -Ve

$-1.0 * 2^{-126} \Rightarrow$

80800000 H Ans

* Denormalized number

- Largest +ve

S E

0 00000000 1111---1

007FFFFF H Ans

exception

$+ \quad 1-2^{-23} * 2^{1-127} \Rightarrow 1-2^{-23} * 2^{-126}$

Exception: when all the exponent are 0's the we'll consider it as "1"

- Smallest ~~+ve~~ -Ve

1 00000000 1111---

807FFFFF H

- Smallest +ve

S E M

0 00000000 0000---1

00000001H Ans $\Rightarrow 2^{-23} * 2^{1-127}$

$\Rightarrow 2^{-23} * 2^{-126}$

$\Rightarrow 2^{-149}$ Ans

- Largest -Ve

1 0000 0000 0000---1

80000001H

$-2^{-23} * 2^{-126} \Rightarrow -2^{-149}$ Ans

Background :-

* Find the smallest diff b/w two Consecutive normalized +ve number

SPN 0 00000001 0000---
0080000OH ⇒ $1.0 * 2^{1-127} = 2^{-126}$

↓ Next Consecutive number

0 00000001 0000---1
$1 + 2^{-23} * 2^{-126}$

gap in Consecutive no. ⇒ $(1 + 2^{-23}) * 2^{-126} - 1 * 2^{-126}$

$2^{-126} + 2^{-149} - 2^{-126}$

* Minimum gap : 2^{-149}

* Max gap :

0 1111 1110 1111---1
7F7FFFFFH ⇒ $2 - 2^{-23} * 2^{127}$

0 1111 1110 0000---0
$1.0 * 2^{254-127} = 2^{127}$

0 1111 1110 000---1
$(1 + 2^{-23}) * 2^{127} \Rightarrow 2^{104}$

gap ⇒ $2^{127} + 2^{104} - 127 \Rightarrow 2^{104}$ Ans

2^{-149} → Min gap b/w normalized number

2^{104} → Max ——— a ——— 11

IEEE → xlon uniform distribution

* gap in denormals

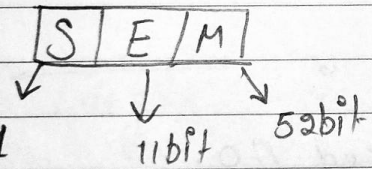
→ to maintain the minimum gap

0 00000000/000 - - - 1

$2^{-23} \times 2^{1-127}$ => 2^{-149}

* IEEE - Notations Part3

• Double Precision



$(1.5)_{10} \Rightarrow (1.1)_2$

0 0 11111111110 - - - 0

3 F F8000000000H Ans

$(.25)_{10} \Rightarrow (.01)_2 \Rightarrow 1.0 \times 2^{-2}$

0 0 11111111101 000 - - - 0

3FD0000000000000H

$(-40.1)_{10}$ → Find Single and double precision

$(101000.00011)_2$ ⇒ $1.0100000\ 00011 \times 2^5$ → Normalized form

Single Precision

1 10000100 01000000011 0110 0110

C 4206666H

* double precision

1 10000000100 0100000 0011 0011 0011 00

C 0440CCCCCCCCCCCDH Ans

Largest Positive Normalized No.

0 11111111110 1111 -- 1

7FEFFFFFFFFFFFFFH

$(2-2^{-52}) * 2^{2046-1023}$

$(2-2^{-52}) \times 2^{1023}$

Smallest ~~positive~~ -ve normalized no.

1 111 1111 0000 1111 ---- 1

FFEFFFFFFFFFFFFH

$-(2-2^{-52}) * 2^{2046-1023}$ =>

$-2 - 2^{-52} * 2^{1023}$

- Smallest postive normalized no.

0 0000000000010000 ----.

$1.0 * 2^{1-1023}$ => 2^{-1022} Ans

- Largest -ve no.

-2^{-1022} Ans

* Denormals

- ~~Largest~~ smallest +ve denormalized no.

0000000000000000 ---- 1

0000000000000001H

$2^{-52} * 2^{1-1023}$ => 2^{-1074} Ans

- Largest -ve denormalized no.

-2^{-1074} Ans

- Largest +ve normalized no.

0 00000000000 1111 --- 1

$(1-2^{-52}) * 2^{1-1023}$

- Smallest -ve denormalized no.

$-(1-2^{-52}) * 2^{-1022}$ Ans

* Max gap b/w two consecutive normalized number

0 1111111110 0000 --- 0

$1.0 \times 2^{2046-1023} \Rightarrow 2^{1023}$

Next Consecutive no.

0 1111111110 000 --- 1

$1.0 * 2^{-52} * 2^{2046-1023}$

$(1+2^{-52})\ 2^{1023}$

$2^{1023} + 2^{971}$

Max gap : 2^{971}

$2^{971} = 10^{x}$

$971 = x * \log_2 10$

$x = 320$

10^{320} Approx

Min gap :

0 00000000001 000 — —

$1.0 * 2^{1-1023} \Rightarrow 2^{-1022}$ Ans

Next Consecutive no

0 00000000010 000 — — 1

$(1.0 + 2^{-52}) * 2^{1-1023}$

$2^{1022} + 2^{-1074}$

Min gap $\Rightarrow$ 2^{-1074} Ans

* Denormals ← min, max gap, same

0 00000000000 0000---1

$2^{-52} * 2^{1-1023} \Rightarrow 2^{-1074}$ Ans

Next Consecutive no.

0 00000000000 0000---10

$2^{-51} * 2^{1-1023} = 2^{-1073}$

gap $2^{-1073} - 2^{-1074}$

$2^{-1073} - 2^{-1} * 2^{-1073}$

$\frac{2^{-1073}}{2} \Rightarrow 2^{-1074}$

Representation of Number System is Completed