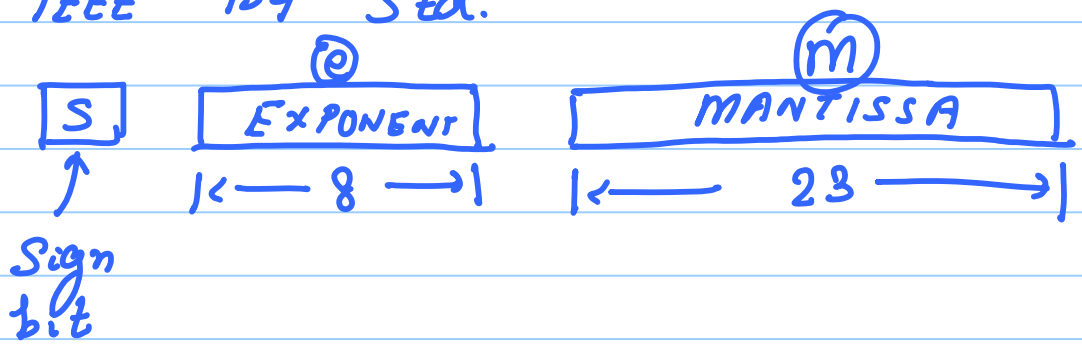
Note Title 05-09-2012

Sep-5.

Represent floating point in binary

$[2.96]_d$ $[5 \times 10^{15}]$ $[-1.923 \times 10^{-27}]$

IEEE 754 32 Std.



$13 = 8 + 4 + 0 + 1$

$2^3 + 2^2 + 0 + 2^0$

1	1	0	1

1.75

$1 + 0.5 + 0.25$

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

Subset of numbers of the form : $1.x$

$1.x$

$1 + 2^{-1}(x_1) + 2^{-2}(x_2) + \cdots\cdots$

Any Number (N) (decimal)

$N = (1.x) \times 2^{exp} \times (\pm 1)$

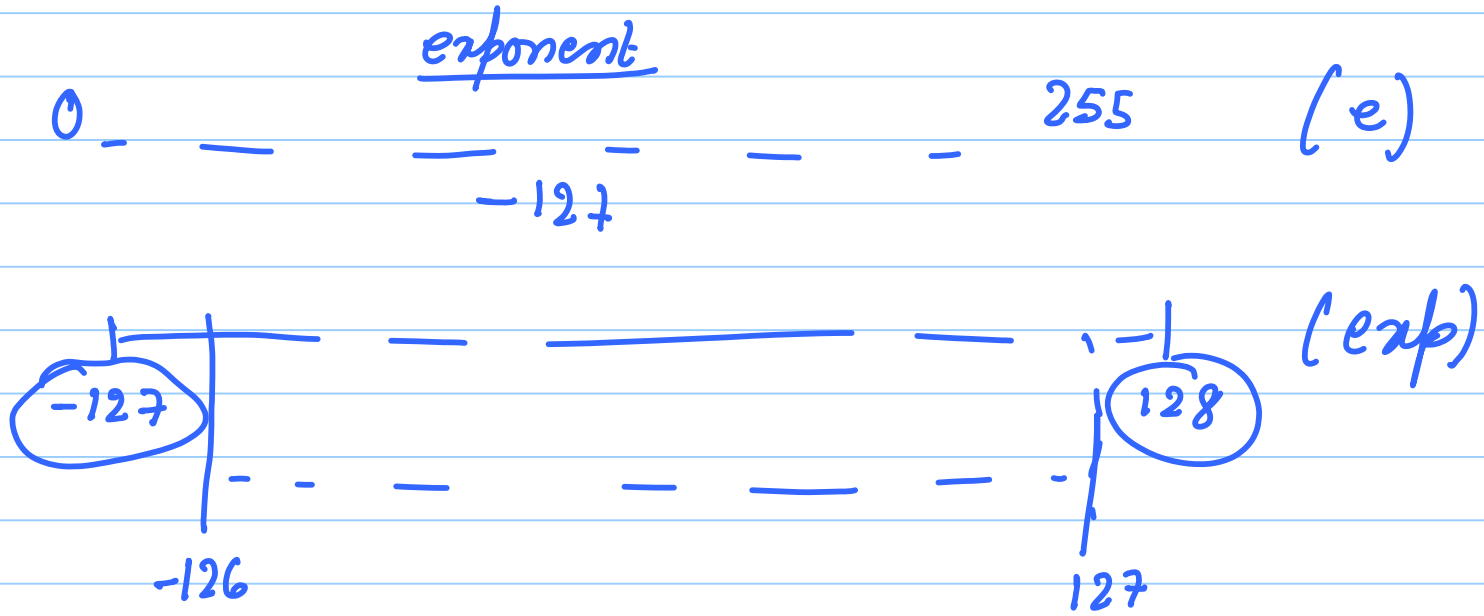
Sign.

Sign ±1 (sign bit) (1)

e → biased notation. [8]

(0, .. 255)

exp = e − BIAS (127)



Mantissa {23 bits}

$$N = (-1)^S \, 1.m \times 2^{(e-127)}$$

Special case:

			N	
	$e = 0$	$m = 0$	0	
$S = 0$,	$e = 255$,	$m = 0$	∞	
$S = 1$,	$e = 255$,	$m = 0$	$-\infty$	
	$e = 255$,	$m \neq 0$	NAN	{NAN + 2 = NAN NAN × 2 = NAN}
	$e = 0$	$m \neq 0$	Denormal Numbers.	

1) Floating point numbers are approximate.

2) There is some amount of error in representation.

$x > 0$, $x/2 > 0$

Normal floating point numbers.

Largest: $\{[1.\overbrace{1 - - - - 1}^{23}]\}_{\text{binary}} \times 2^{127}$

(+ve)

(Normal)

$\Downarrow$

$(1 + 2^{-1} - - - - + 2^{-23})$

$(2 - 2^{-23}) \times 2^{127}$

$= 2^{128} - 2^{104}$

Smallest: $1.\underbrace{0\cdots0}_{23} \times 2^{-126} = 2^{-126}$
(+ve)
(Normal)

Prog.
```
x = 2^-126
if (x/2 > 0)
    print ("hi");
```

Denormal number -

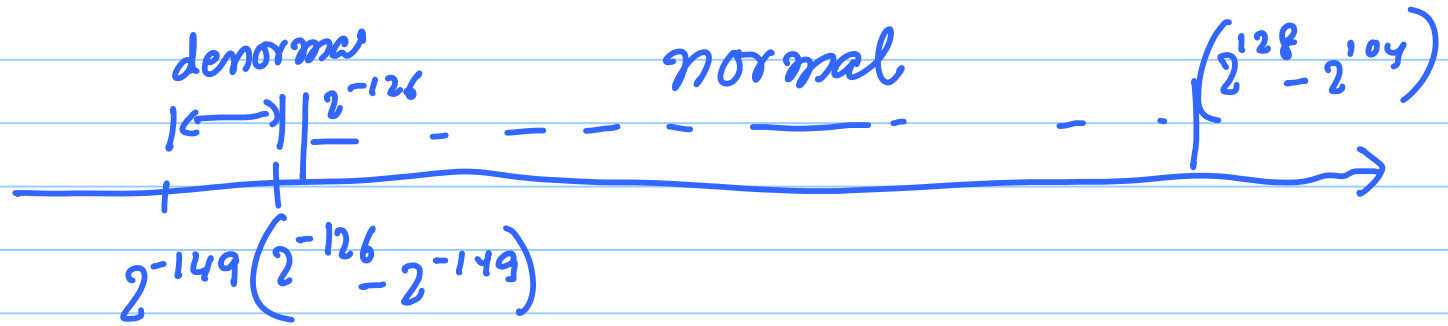
$e = 0$ $m \neq 0$

Form: $(-1)^s \; 0.m \times 2^{-126}$

Largest denormal number:
(+ve)

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

Smallest (+ve) denormal number: 2^{-149}



Is, $(\Delta x > 0)$

$x + \Delta x > x$

2^{100} 2^{-100}

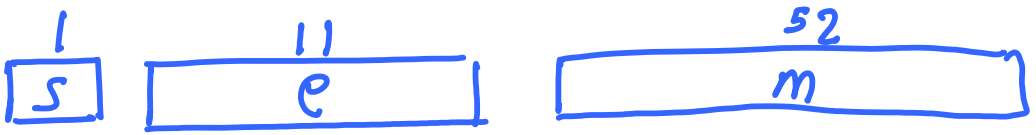
```
printf ( x+Δx-x );
```

$x = 2^{-149}$

```
if (x/2 >0)
    printf ("hi");
```

float → 32 bits double → 64 bits

double.



$exp = e - 1023$