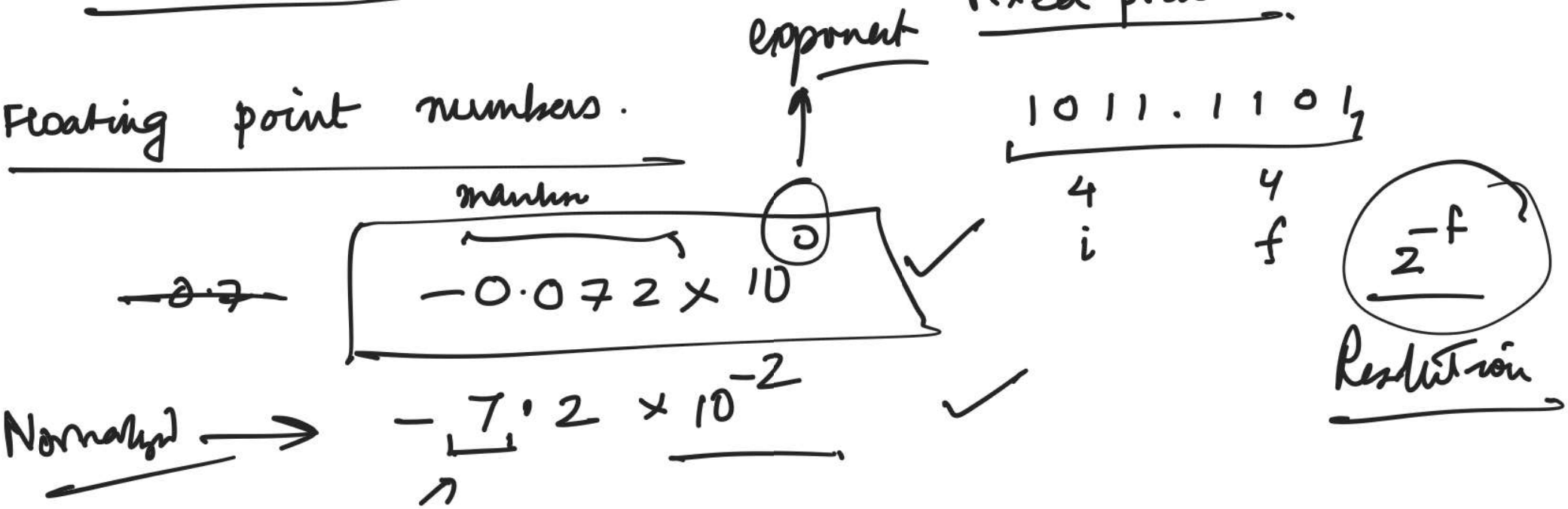
IEEE 754

Floating point numbers.



Fixed point c.

1011.1101₂

4 4

i f

2^{-f}

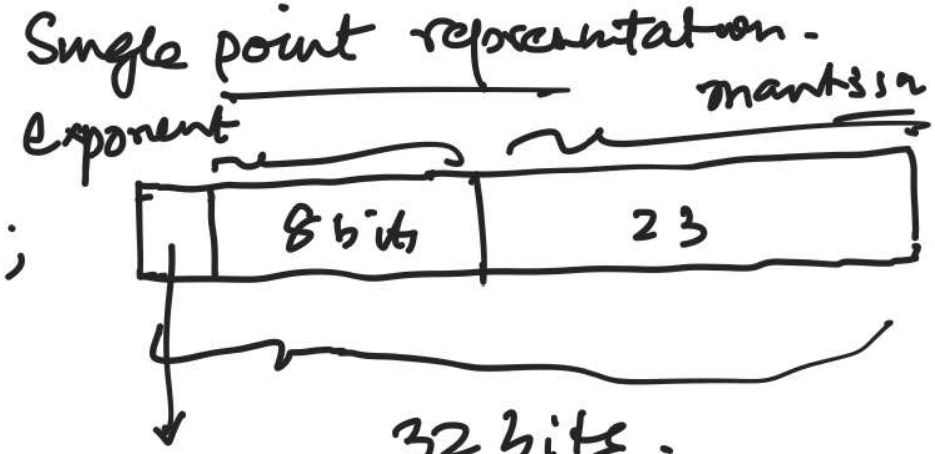
Resolution

IEEE 754. 1985.

19.59375_{10}

float a;

Single point representation.

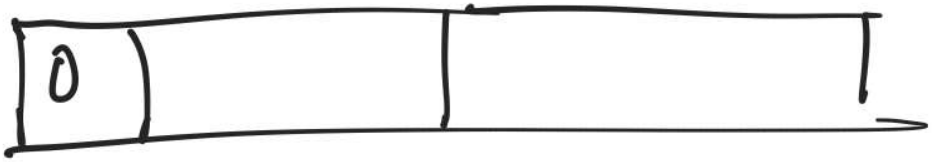


(1) Determine the sign bit.

Sign bit → 0

(for negative number → 1)

Sign



(2)

2 | 19	
2 | 9	1
2 | 4	1
2 | 2	0
2 | 1	0
0 | 1	1

$0.59375 \times 2 = 1.1875$ → 1

$0.1875 \times 2 = 0.375$ → 0

$0.375 \times 2 = 0.75$ → 0

$0.75 \times 2 = 1.5$ → 1

$0.5 \times 2 = 1.0$ → 1

Binary representation: 10011.10011

i = 5, f = 5

Step 3 Normalize

(place the binary point after leftmost 1)

$1.00111\,0011 \times 2^{4}$

4 → exponent.

mantissa.

Step 4 Biased Exponent.

Unbiased exponent : 4.

Biased exponent : 127 + 4 = 131

Bias/offset $[2^{n-1} - 1]$

$n = 8$

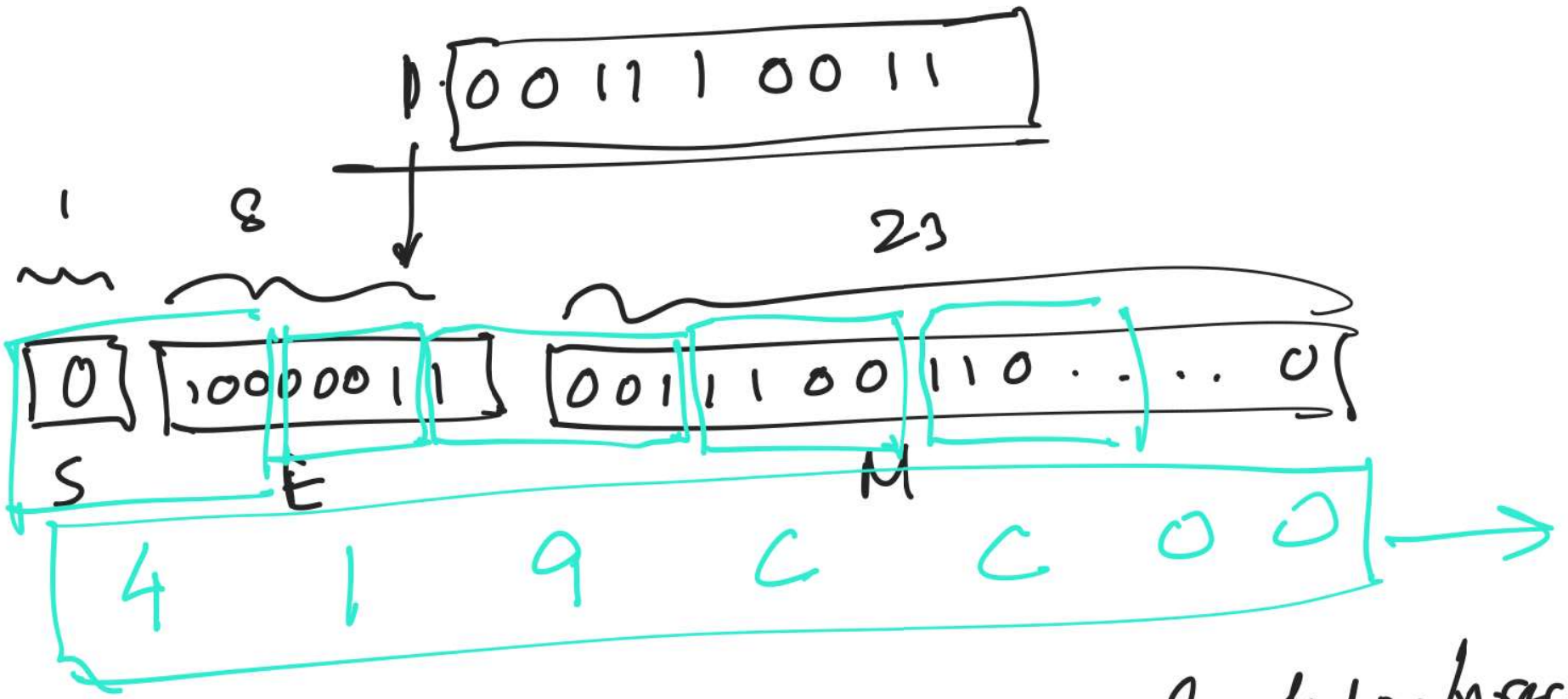
$[2^7 - 1]$

-4

8 bits

1	0	0	0	0	0	1	1

Step 5 → Mantissa →



No. of bits in mantissa → Resolution/precision.

" " " " exponent → range.

0.375

0.375_{10}

0.011×2^{0}

1.1×2^{-2}

Unbiased Exponent = -2.

Biased Exponent = 127 - 2 = 125.

Sign : 0

ignore.

3EC00000

0	01111101	10.....0.

Example

C8500000

1 | 100 1000 0 | 101 0000 0000 0000

Sign

Unbiased exponent

144

144 − 127

(1.101000 0 0)

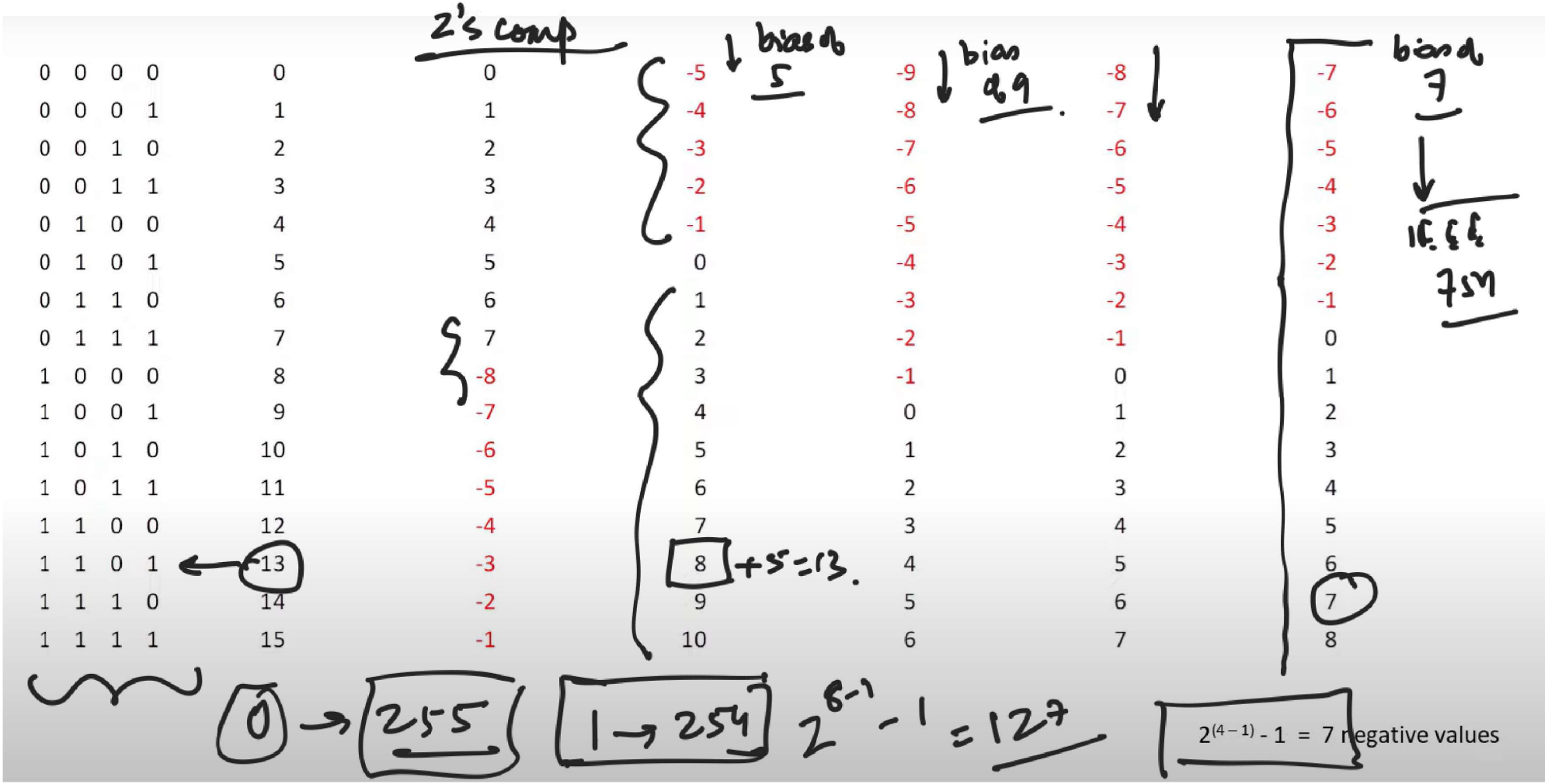
mantissa

-1.101×2^{17}

-1.625×2^{17}

Next 3 slides:

Ack:
https://www.youtube.com/watch?v=RuKkePyo9zk&list=PLTd6ceoshprcpen2Jvs_JiuvWvqIAkzea&index=10


2's comp
biased 5
bias 9
bias 7
0 0 0 0 0 0 -5 -9 -8 -7
0 0 0 1 1 1 -4 -8 -7 -6
0 0 1 0 2 2 -3 -7 -6 -5
0 0 1 1 3 3 -2 -6 -5 -4
0 1 0 0 4 4 -1 -5 -4 -3
0 1 0 1 5 5 0 -4 -3 -2
0 1 1 0 6 6 1 -3 -2 -1
0 1 1 1 7 7 2 -2 -1 0
1 0 0 0 8 -8 3 -1 0 1
1 0 0 1 9 -7 4 0 1 2
1 0 1 0 10 -6 5 1 2 3
1 0 1 1 11 -5 6 2 3 4
1 1 0 0 12 -4 7 3 4 5
1 1 0 1 13 -3 8 +5=13. 4 5 6
1 1 1 0 14 -2 9 5 6 7
1 1 1 1 15 -1 10 6 7 8
0 → 255
1 → 254
2^(8-1) - 1 = 127
2^(4-1) - 1 = 7 negative values

IEEE 754 Format	Sign	Exponent	Mantissa	Exponent Bias
32 bit single precision	1 bit	8 bits	23 bits (+ 1 not stored)	$2^{(8-1)} - 1 = 127$
64 bit double precision	1 bit	11 bits	52 bits (+ 1 not stored)	$2^{(11-1)} - 1 = 1023$
128 bit quadruple precision	1 bit	15 bits	112 bits (+ 1 not stored)	$2^{(15-1)} - 1 = 16383$

double a;

Reserved Exponent Values

Exponent Value	Mantissa	Represents
✓ 11111111	All zeros	Infinity (∞)
✓ 11111111	Not all zeros	Not a number (NAN)
00000000	All zeros	Zero
00000000	Not all zeros	Subnormal (very small)

Practice problems

Convert the following decimal numbers to 32-bit IEEE 754 format by hand:

a. 1.0
b. −0.1
c. 2016.0
d. 0.00390625
e. −3125.3125
f. 0.33
g. −0.67
h. 3.14

▼ Answer

a. `3f800000`
b. `bdcccccd`
c. `44fc0000`
d. `3b800000`
e. `c5435500`
f. `3ea8f5c3`
g. `bf2b851f`
h. `4048f5c3`

Convert the following hexadecimal numbers to decimal by hand using the 32-bit IEEE 754 format:

a. `40000000`
b. `bf800000`
c. `3d800000`
d. `c1804000`
e. `42c81000`
f. `3f99999a`
g. `42f6e666`
h. `c25948b4`

▼ Answer

a. +2.0
b. −1.0
c. +0.0625
d. −16.03125
e. 100.03125
f. 1.2
g. 123.449997
h. −54.320999

Ack: https://bob.cs.sonoma.edu/IntroCompOrg-RPi/exercises-32.html