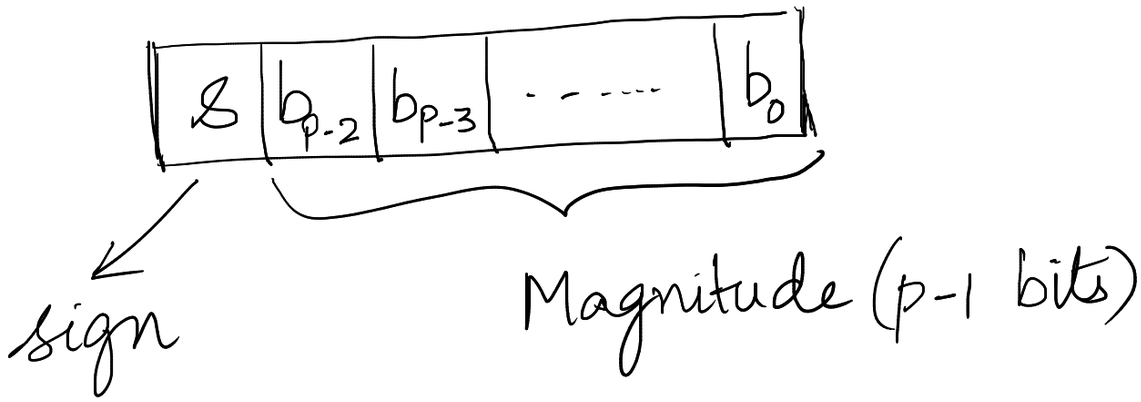
# Number Representation in Computers

Aditya Tatu

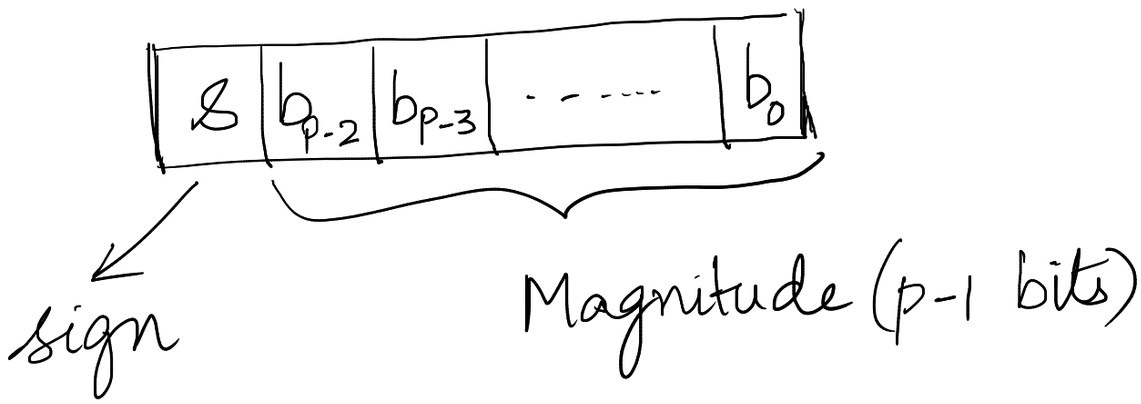
# INTEGERS

* Assume p bits:

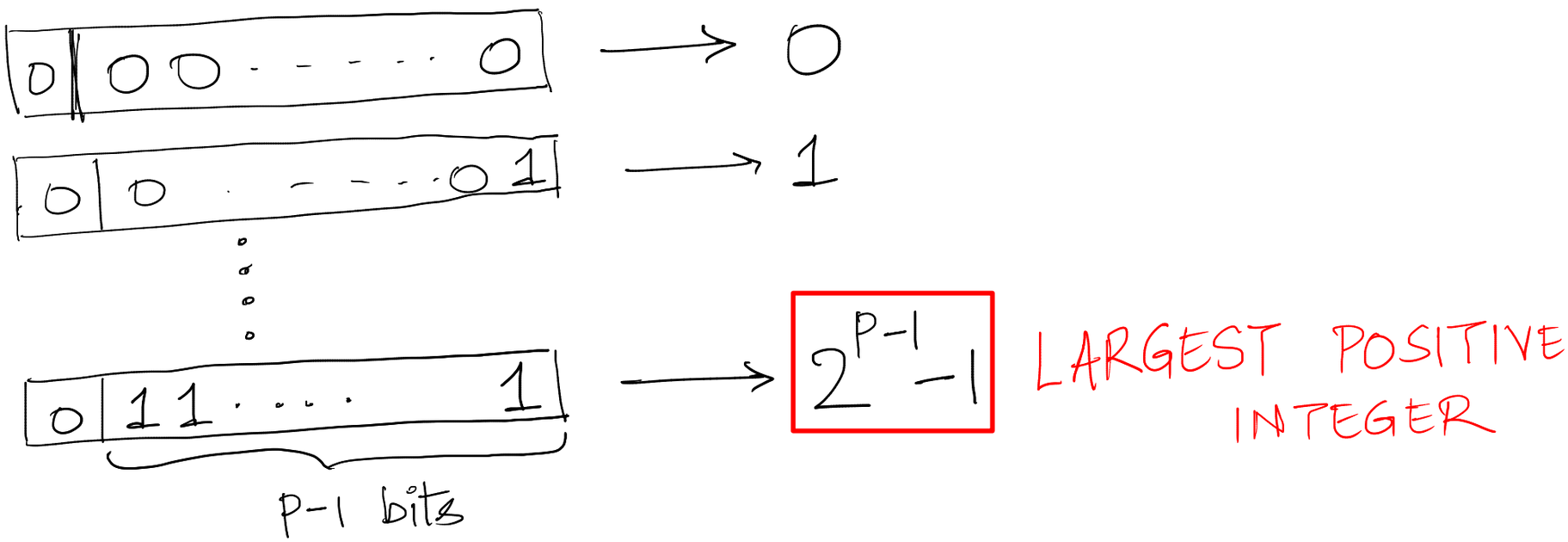


# INTEGERS

* Assume $p$ bits:

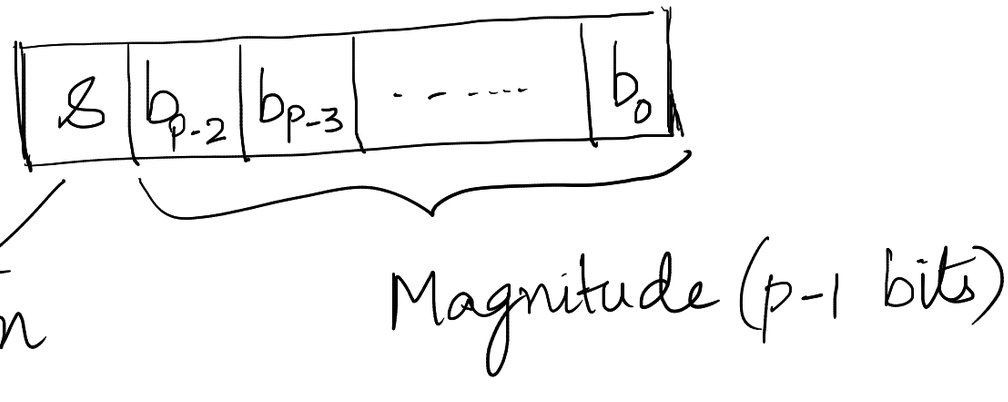



* **POSITIVE INTEGERS**: $s = 0$



# INTEGERS

* Assume p bits:




## NEGATIVE INTEGERS:

* Use 2s complement representation
  - → 1s complement + 1
  - → Starting from LSB, replicate bits till you encounter a 1, replicate this 1, after which all bits are flipped.

# NEGATIVE INTEGERS:

## Example of 2s complement:

ⓐ $10110 \xrightarrow{2s} 01010$

ⓑ $1000 \xrightarrow{2s} 1000$

## * Negative Integer representation using 2s complement

Let $z < 0$. $\boxed{Rep(z) = 2s\,comp(|z|_2)}$

Example: ⓐ $-9$: $(|9|_2) = \underbrace{0}_{sign}1001$

$$2s\,comp(01001) = \underbrace{1}_{sign}0111$$

$$\therefore Rep(-9) = \boxed{1\,|\,0111}$$

② $z = -18.$ $\quad (|-18|_2) = \underbrace{0}_{\text{sign}}\,|\,\underbrace{010010}_{18}$

(p=7 bits)

$$2_s\ \text{Comp}(0010010) = \boxed{\underbrace{1}_{\text{sign}}\,|\,101110}$$

---

Q. What is the largest magnitude negative number representable using p-bits (incl. sign bit)

② $z = -18.$ $\quad (|-18|_2) = \underbrace{0}_{\text{sign}}|\underbrace{010010}_{18}$

(p=7 bits)

$$2s\ Comp(0010010) = \boxed{\underbrace{1}_{\text{sign}}|101110}$$

---

Q. What is the largest magnitude negative number representable using p-bits (incl. sign bit)

$Rep(-1) = 2s\ Comp(1_2) = 2s\ comp(0|0\cdots01) = 1|11\cdots1$

(All 1's)

$Rep(-2) = 2s\ Comp(2_2) = 2s\ comp(0|0\cdots10) = 1|1\cdots10$

② $z = -18.$ $\quad (|-18|_2) = \underbrace{0}_{\text{sign}}\,|\,\underbrace{010010}_{18}$

(p=7 bits)

$$2s\ Comp(0010010) = \boxed{\underbrace{1}_{\text{sign}}\,|\,101110}$$

---

Q. What is the largest magnitude negative number representable using p-bits (incl. sign bit)

$Rep(-1) = 2s\ Comp(1_2) = 2s\ comp(0|0\cdots01) = 1|11\cdots1$

(All 1's)

$Rep(-2) = 2s\ Comp(2_2) = 2s\ comp(0|0\cdots10) = 1|1\cdots10$

Q. What is the number represented by the bits $\boxed{1\,|\,\underbrace{00\cdots\cdots01}_{p-1}}$

sign

② $z = -18.$ $\quad (|-18|_2) = \underbrace{0}_{\text{sign}}|\underbrace{010010}_{18}$

(p=7 bits)

$$2s\ Comp(0010010) = \boxed{\underbrace{1}_{\text{sign}}|101110}$$

---

Q. What is the largest magnitude negative number representable using p-bits (incl. sign bit)

$$Rep(-1) = 2s\ Comp(1_2) = 2s\ comp(0|0\cdots01) = 1|\underbrace{11\cdots1}_{\text{(All 1's)}}$$

$$Rep(-2) = 2s\ Comp(2_2) = 2s\ comp(0|0\cdots10) = 1|1\cdots10$$

Q. What is the number represented by the bits $\boxed{\underbrace{1}_{\text{sign}}|\underbrace{00\cdots\cdots01}_{p-1}}$ $\longrightarrow$ sign $= 1 \Rightarrow z < 0.$

$$|z| = 2s\ Comp(100\cdots01)$$

$$= \underbrace{0}_{\text{sign}}|\underbrace{111\cdots1}_{p-1} = 2^{p-1} - 1$$

Q. What is the largest magnitude negative number representable using p-bits (incl. sign bit)

$$Rep(-1) = 2s\,Comp(1_2) = 2s\,comp(0|0\cdots01) = 1|11\cdots1$$

(All 1's)

$$Rep(-2) = 2s\,Comp(2_2) = 2s\,comp(0|0\cdots10) = 1|1\cdots10$$

Q. What is the number represented by the bits $\boxed{1|\underbrace{00\cdots\cdots01}_{p-1}}$ (sign) $\longrightarrow$ sign $= 1 \Rightarrow z < 0$.

$$|z| = 2s\,Comp(100\cdots01)$$

$$= \underbrace{0}_{sign}|\underbrace{111\cdots1}_{p-1} = 2^{p-1}-1$$

Similarly, the bit seq $\boxed{1|\underbrace{00\cdots\cdots0}_{p-1}}$ represents $z < 0$

such that $|z| = 2s\,comp(10\cdots0) = 10\cdots0 = 2^{p-1}$

$$\Rightarrow \boxed{z = -2^{p-1}}$$

# SUMMARY

1. With $p$ bits (incl. sign bit), largest positive integer $= 2^{p-1} - 1$

   Bit sequence: $\underbrace{0}_{\text{sign}}\underbrace{11\cdots1}_{p-1}$

2. With $p$ bits (incl. sign bits), negative integer with largest magnitude $= -2^{p-1}$. Bit seq: $\underbrace{1}_{\text{sign}}\underbrace{00\cdots0}_{p-1}$

RANGE OF INTEGERS with $p$-bits: $\left[-2^{p-1}, 2^{p-1}-1\right]$

# FLOATING POINT REPRESENTATION OF RATIONALS

* Fixed point representation : (3 bits for Integer part, 4 for fractional)

$$(5.25)_2 = 101.0100$$

* Floating point representation : (3 bits for exponents, 4 for fractional)

$$(5.25)_2 = 101.0100 \rightarrow \underbrace{\boxed{1}}_{\text{always 1}}.\underbrace{\boxed{0101}}_{\text{fraction}} \times 2^{\overbrace{\boxed{2}}^{\text{exponent}}}$$

| | | |
|---|---|---|
| 1 | 0 | 1 |

Exponent
(excess-3)

| | | | |
|---|---|---|---|
| 0 | 1 | 0 | 1 |

Fraction

# FLOATING POINT REPRESENTATION OF RATIONALS

* Fixed point representation : (3 bits for Integer part, 4 for fractional)

$$(5.25)_2 = 101.0100$$

* Floating point representation : (3 bits for exponents, 4 for fractional)

$$(5.25)_2 = 101.0100 \rightarrow \boxed{1}.\boxed{0101} \times 2^{\boxed{2}}$$

always 1 fraction exponent

| | | |
|---|---|---|
| 1 | 0 | 1 |

Exponent (excess-3)

| | | | |
|---|---|---|---|
| 0 | 1 | 0 | 1 |

Fraction

* Range for

a. Fixed point (Ignoring negatives) : $[0, 8)$
   - Uniformly sampled with interval $2^{-4} = 1/16$

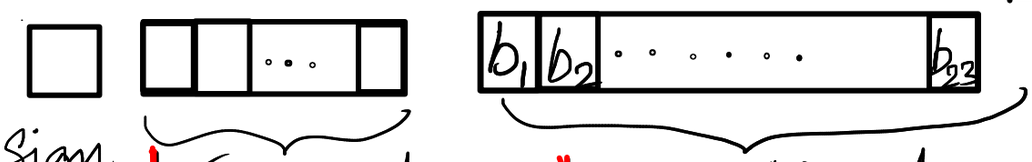
b. Floating point (Ignoring negatives) : $[0, 16)$
   - Non-uniformly sampled; interval ranges from $1/64, \ldots, 1/2$.

# IEEE Floating Point Representation Standard

## * Single Precision : 32 bits

$\boxed{b_0}.b_1 b_2 \cdots b_{23} \times 2^e$

- → assumed as 1 → Normal floats
- → If $\neq 1$ (De/Sub-normal floats)




Sign
0 → +ve
1 → -ve

* Exponent 8 bits
* Excess-127
* $0 \le e' \le 255$
* Actual range: $[-126, 127]$

$e' \in \{0, 255\}$ reserved

* Fractional part - 23 bits
* Typically $b_0 = 1$

## * Double Precision : 64 bits

$\boxed{b_0}.b_1 b_2 \cdots b_{52} \times 2^e$

- → assumed as 1 → Normal floats
- → If $\neq 1$, (De/Subnormal floats)

Sign
0 → +ve
1 → -ve

* Exponent 11 bits
* Excess-1023
* $0 \le e' \le 2047$
* Actual Range: $[-1022, 1023]$
* $e' \in \{0, 1047\}$ reserved

* Fractional part - 52 bits
* Typically $b_0 = 1$

# Special Numbers (Both Single & Double Precision)

| Sign bit | Exponent bits | Mantissa bits | Interpretation |
|---|---|---|---|
| 0 | All 0's | All 0's | $+0.0$ |
| 1 | All 0's | All 0's | $-0.0$ |

If all Exponent bits are 0's, <u>AND</u> some Mantissa bits are non-zero, it indicates a De-/Sub-normal number

$\Rightarrow \boxed{b_0 = 0}$

| | | | |
|---|---|---|---|
| 0 or 1 | All 0's | $b_1 b_2 \cdots b_{23}$ (Atleast 1 bit $\neq 0$) | $N_0 = \pm 0.b_1 b_2 \cdots b_{23} \times 2^{-126}$ |
| | | $b_1 b_2 \cdots b_{52}$ (Atleast 1 bit $\neq 0$ | $N_0 = \pm 0.b_1 b_2 \cdots b_{52} \times 2^{-1022}$ |

| * Smallest Sub-normal positive no. | | * Largest Positive Subnormal no. | |
|---|---|---|---|
| • $SP = 2^{-23} \times 2^{-126} = 2^{-149}$ | • $DP = 2^{-52} \times 2^{-1022} = 2^{-1074}$ | • $SP = 1 - (1/2)^{23}$ | • $DP = 1 - (1/2)^{52}$ |

# Special Numbers (SP & DP)

| Sign bit | Exponent | Mantissa | Interpretation |
|---|---|---|---|
| 0 or 1 | All 1's | All 0's | $\pm\infty$ |
| 0 or 1 | All 1's | Non-zero | NAN (Not a number)<br>$0/0$, $0\times\infty$ |

# S.P. NORMALIZED floats (Exponent $\notin \{(00\cdots0);(11\cdots1)\}$)

| Sign | Exponent | Mantissa | Interpretation |
|---|---|---|---|
| 0 | $0\cdots01_8$ | $00\cdots0_{23}$ | Exponent $e' = 1$, $e = 1-127 = -126$<br>Also $b_0 = 1$.<br>$\therefore N_0 = 1.0 \times 2^{-126} = 2^{-126}$ |
| 0 | $0\cdots01_8$ | $0\cdots01_{23}$ | $e = -126$, $b_0 = 1$<br>$N_0 = 2^{-126} + 2^{-149}$ |
| 0 | $0\cdots01_8$ | $0\cdots10_{23}$ | $e = -126$, $N_0 = 2^{-126} + 2^{-148}$ |
| 0 | $0\cdots01_8$ | $11\cdots11_{23}$ | $e = -126$, $N_0 = 2^{-126} + 2^{-127} + 2^{-128} + \cdots + 2^{-149}$<br>$= 2^{-125}(1-2^{-23})$ |
| 0 | $0\cdots10_8$ | $00\cdots0_{23}$ | $e = -125$, $N_0 = 2^{-125}$ |
| 0 | $0\cdots10_8$ | $00\cdots01_{23}$ | $e = -125$, $N_0 = 2^{-125} + 2^{-148}$ |
| 0 | $0\cdots10_8$ | $11\cdots11_{23}$ | $e = -125$, $N_0 = 2^{-124}(1-2^{-23})$ |

# S.P. NORMAL floats (Exponent $\notin \{(0\,0\cdots0);(1\,1\cdots1)\}$)

| Sign | Exponent | Mantissa | Interpretation |
|---|---|---|---|
| 0 | $0\cdots01_8$ | $00\cdots0_{23}$ | Exponent $e' = 1$, $e = 1-127 = -126$<br>Also $b_0 = 1$.<br>$\therefore N_0 = 1.0 \times 2^{-126} = 2^{-126}$ |
| 0 | $0\cdots01_8$ | $0\cdots01_{23}$<br>gap between numbers $= 2^{-149}$ | $e = -126$, $b_0 = 1$<br>$N_0 = 2^{-126} + 2^{-149}$ |
| 0 | $0\cdots01_8$ | $0\cdots10_{23}$ | $e = -126$, $N_0 = 2^{-126} + 2^{-148}$ |
| 0 | $0\cdots01_8$ | $11\cdots11_{23}$ | $e = -126$, $N_0 = 2^{-126} + 2^{-127} + 2^{-128} + \cdots + 2^{-149}$<br>$= 2^{-125}(1-2^{-23})$ |
| 0 | $0\cdots10_8$ | $00\cdots0_{23}$ | $e = -125$, $N_0 = 2^{-125}$ |
| 0 | $0\cdots10_8$ | $00\cdots01_{23}$<br>gap bet$^n$ numbers $= 2^{-148}$ | $e = -125$, $N_0 = 2^{-125} + 2^{-148}$ |
| 0 | $0\cdots10_8$ | $11\cdots11_{23}$ | $e = -125$, $N_0 = 2^{-124}(1-2^{-23})$ |

# S.P. NORMALIZED Floats

| Sign | Exponent | Mantissa | Interpretation |
|---|---|---|---|
| 0 | $011\cdots1_8$ | $00\cdots0_{23}$ | $e'=127,\ e=0,\ N_0=1.0\times2^0=1.0$ |
| 0 | $011\cdots1_8$ | $00\cdots01_{23}$ | $e=0,\ \underline{N_0=1+2^{-23}}$ |

* The difference between 1 and the next floating point number is called the Machine precision (eps)

For Single Precision $eps=2^{-23}$

For Double Precision $eps=2^{-53}$ [Gap

| | | | |
|---|---|---|---|
| 0 | $11\cdots10_8$ | $00\cdots0_{23}$ | $e'=254,\ e=127,\ N_0=1.0\times2^{127}$ |
| 0 | $11\cdots10_8$ | $00\cdots01_{23}$ | $e=127,\ N_0=2^{127}+2^{104}$ |
| 0 | $11\cdots10_8$ | $00\cdots010_{23}$ | $e=127,\ N_0=2^{127}+2^{105}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ |
| 0 | $11\cdots10_8$ | $11\cdots11_{23}$ | $e=127,\ N_0=2^{104}(2^{24}-1)$ |

# S.P. NORMALIZED Floats

| Sign | Exponent | Mantissa | Interpretation |
|---|---|---|---|
| 0 | $011\cdots1_8$ | $00\cdots0_{23}$ | $e'=127,\ e=0,\ N_0=1.0\times2^0=1.0$ |
| 0 | $011\cdots1_8$ | $00\cdots01_{23}$ | $e=0,\ \underline{N_0=1+2^{-23}}$ |

* The difference between 1 and the next floating point number is called the Machine precision (eps)

For Single Precision $eps=2^{-23}$

For Double Precision $eps=2^{-53}$

[Gap between $[1,2)$ $=eps$]

| | | | |
|---|---|---|---|
| 0 | $11\cdots10_8$ | $00\cdots0_{23}$ | $e'=254,\ e=127,\ N_0=1.0\times2^{127}$ |
| 0 | $11\cdots10_8$ | $00\cdots01_{23}$ | $e=127,\ N_0=2^{127}+2^{104}$ |
| 0 | $11\cdots10_8$ | $00\cdots010_{23}$ | $e=127,\ N_0=2^{127}+2^{105}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ |
| 0 | $11\cdots10_8$ | $11\cdots11_{23}$ | $e=127,\ N_0=\boxed{2^{104}(2^{24}-1)}$ |

LARGEST POSITIVE FLOAT