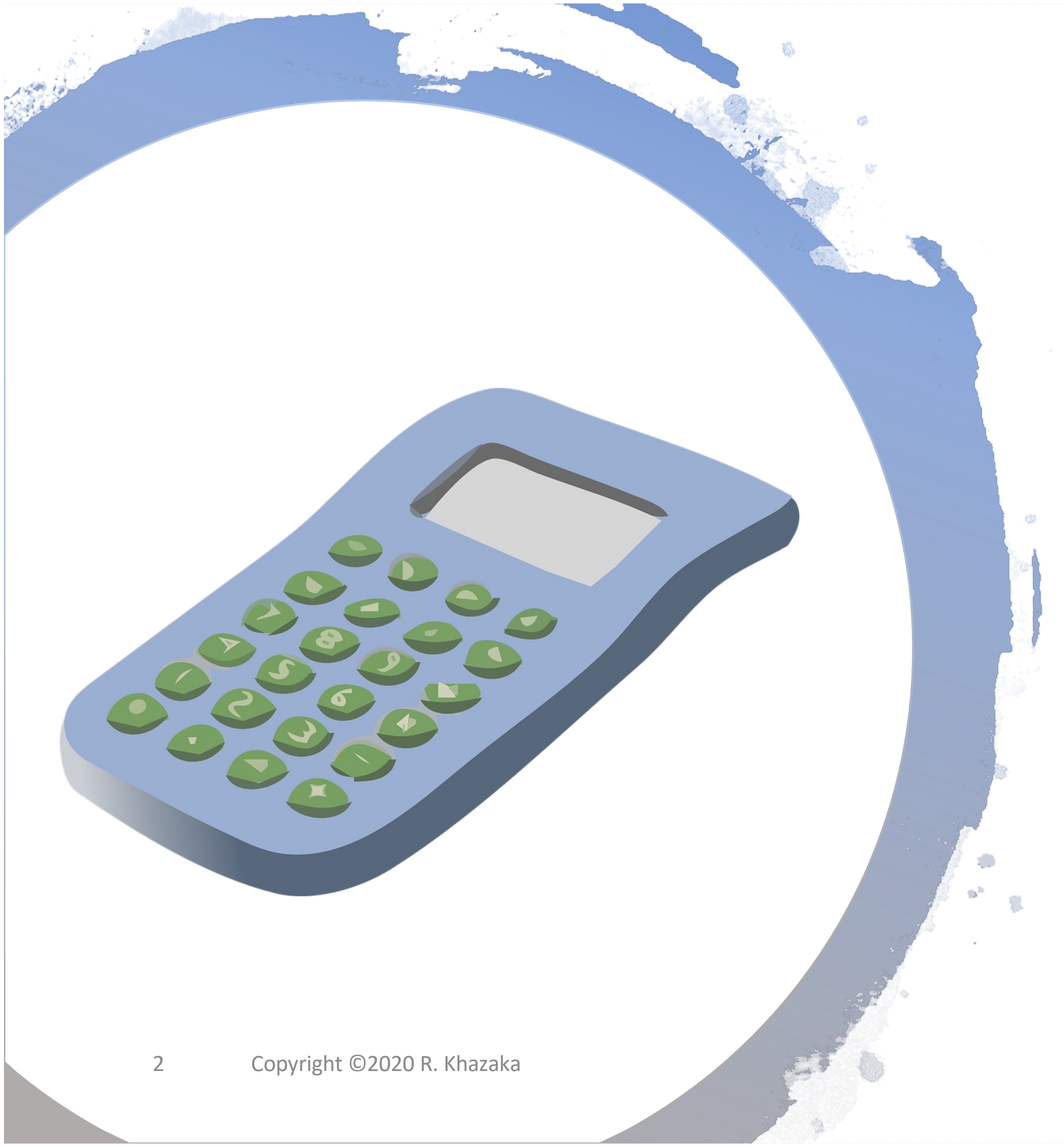
ECSE 343 Numerical Methods in Engineering

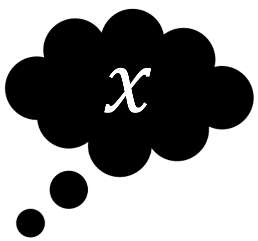
Roni Khazaka
Dept. of Electrical and Computer Engineering
McGill University



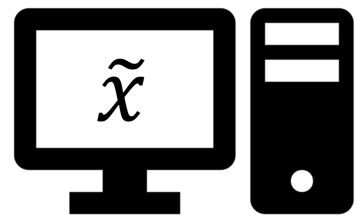
Number Representation

Number Representation

Abstract Quantity



Stored Quantity



$$7x = 4 \xrightarrow{x = 4/7} \tilde{x} \begin{cases} 0.6 \\ 0.57 \\ 0.57143 \\ 0.571429 \\ 0.57142857 \end{cases}$$

Unsigned Integers

10^3	10^2	10^1	10^0
8	3	0	2

$\rightarrow$ $8\times10^3 + 3\times10^2 + 0\times10^1 + 2\times10^0 = 8{,}302$

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
1	0	1	0	0	1	0	1

8 bit binary

$\rightarrow$ $1\times2^7 + 0\times2^6 + 1\times2^5 + 0\times2^4 + \cdots$

$\cdots + 0\times2^3 + 1\times2^2 + 0\times2^1 + 1\times2^0 = 165_{10}$

d_7	d_6	d_5	d_4	d_3	d_2	d_1	d_0

$\rightarrow$ $\sum_{i=0}^{7} d_i 2^i$

Range: 0 – 255

Unsigned Integers

$$\text{uint8(5)} \rightarrow \underbrace{\begin{array}{|c|c|c|c|c|c|c|c|}\hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ \hline \end{array}}_{\text{8 bits}}$$

uint8() has a range from <u>0 to 255</u>

$$\text{uint16(5)} \rightarrow \underbrace{\begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}\hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ \hline \end{array}}_{\text{16 bits}}$$

uint16() has a range from <u>0 to 65,535</u>

Signed Integers: Sign Bit

±	2^6	2^5	2^4	2^3	2^2	2^1	2^0
s	d_6	d_5	d_4	d_3	d_2	d_1	d_0

$\underbrace{\hphantom{s\ d_6\ d_5\ d_4\ d_3\ d_2\ d_1\ d_0}}_{\text{8 bit binary}}$

$$\rightarrow \quad (-1)^s \sum_{i=0}^{6} d_i 2^i$$

Range: $-127 \rightarrow +127$

±	2^6	2^5	2^4	2^3	2^2	2^1	2^0
1	0	1	0	0	1	0	1

$$\rightarrow \quad -(0\times2^6 + 1\times2^5 + 0\times2^4 + 0\times2^3 + 1\times2^2 + 0\times2^1 + 1\times2^0) = \\ = -37$$

0	0	0	0	0	0	0	0

$\rightarrow \quad +0$

1	0	0	0	0	0	0	0

$\rightarrow \quad -0$

Sign bit: 4-bit example

Unsigned Value (d)	Stored Bits	Signed Value
0	0000	+0
1	0001	+1
2	0010	+2
3	0011	+3
4	0100	+4
5	0101	+5
6	0110	+6
7	0111	+7

Unsigned Value (d)	Stored Bits	Signed Value
8	1000	−0
9	1001	−1
10	1010	−2
11	1011	−3
12	1100	−4
13	1101	−5
14	1110	−6
15	1111	−7

One’s complement

Unsigned Value (d)	Stored Bits	Signed Value
0	0000	+0
1	0001	+1
2	0010	+2
3	0011	+3
4	0100	+4
5	0101	+5
6	0110	+6
7	0111	+7

	Unsigned Value (d)	Stored Bits	Signed Value
$7+8=2^4-1$	8	1000	−7
$6+9=2^4-1$	9	1001	−6
$5+10=2^4-1$	10	1010	−5
$4+11=2^4-1$	11	1011	−4
$3+12=2^4-1$	12	1100	−3
$2+13=2^4-1$	13	1101	−2
$1+14=2^4-1$	14	1110	−1
$0+15=2^4-1$	15	1111	−0

One's complement

Unsigned Value (d)	Stored Bits	Signed Value
0	0000	+0
1	0001	+1
2	0010	+2
3	0011	+3
4	0100	+4
5	0101	+5
6	0110	+6
7	0111	+7

	Unsigned Value (d)	Stored Bits	Signed Value
$7+8=2^4-1$	8	1000	−7
$6+9=2^4-1$	9	1001	−6
$5+10=2^4-1$	10	1010	−5
$4+11=2^4-1$	11	1011	−4
$3+12=2^4-1$	12	1100	−3
$2+13=2^4-1$	13	1101	−2
$1+14=2^4-1$	14	1110	−1
$0+15=2^4-1$	15	1111	−0

$$
\begin{array}{r}
0011 \rightarrow 3_{10} \\
+\ 1010 \rightarrow -5_{10} \\
\hline
1101 \rightarrow -2_{10}
\end{array}
$$

$$
\begin{array}{r}
0111 \rightarrow 7_{10} \\
+\ 1010 \rightarrow -5_{10} \\
\hline
10001 \rightarrow +1_{10}
\end{array}
$$

$$
\begin{array}{r}
1101 \rightarrow -2_{10} \\
+\ 1100 \rightarrow -3_{10} \\
\hline
11001 \rightarrow -6_{10}
\end{array}
$$

Two’s complement

Unsigned Value (d)	Stored Bits	Signed Value
0	0000	0
1	0001	+1
2	0010	+2
3	0011	+3
4	0100	+4
5	0101	+5
6	0110	+6
7	0111	+7

	Unsigned Value (d)	Stored Bits	Signed Value
$8+8=2^4$	8	1000	−8
$7+9=2^4$	9	1001	−7
$6+10=2^4$	10	1010	−6
$5+11=2^4$	11	1011	−5
$4+12=2^4$	12	1100	−4
$3+13=2^4$	13	1101	−3
$2+14=2^4$	14	1110	−2
$1+15=2^4$	15	1111	−1

Two's complement

Unsigned Value (d)	Stored Bits	Signed Value
0	0000	0
1	0001	+1
2	0010	+2
3	0011	+3
4	0100	+4
5	0101	+5
6	0110	+6
7	0111	+7

	Unsigned Value (d)	Stored Bits	Signed Value
$8 - 2^4 = -8$	8	1000	−8
$9 - 2^4 = -7$	9	1001	−7
$10 - 2^4 = -6$	10	1010	−6
$11 - 2^4 = -5$	11	1011	−5
$12 - 2^4 = -4$	12	1100	−4
$13 - 2^4 = -3$	13	1101	−3
$14 - 2^4 = -2$	14	1110	−2
$15 - 2^4 = -1$	15	1111	−1

$$
\begin{array}{r}
0011 \rightarrow 3_{10} \\
+\ 1010 \rightarrow -6_{10} \\
\hline
1101 \rightarrow -6_{10}
\end{array}
$$

$$
\begin{array}{r}
0111 \rightarrow 7_{10} \\
+\ 1010 \rightarrow -6_{10} \\
\hline
10001 \rightarrow +1_{10}
\end{array}
$$

$$
\begin{array}{r}
1101 \rightarrow -3_{10} \\
+\ 1100 \rightarrow -4_{10} \\
\hline
11001 \rightarrow -7_{10}
\end{array}
$$

Two's Complement

2^6	2^5	2^4	2^3	2^2	2^1	2^0

0	1	0	1	1	0	1	1

$\rightarrow$ $1\times2^6+0\times2^5+1\times2^4+1\times2^3+0\times2^2+1\times2^1+1\times2^0=91$

$\downarrow$ Flip all bits and add 1

1	0	1	0	0	1	0	1

$\rightarrow$ $1\times2^7+0\times2^6+1\times2^5+0\times2^4+0\times2^3+1\times2^2+0\times2^1+1\times2^0$

$$=165_{unsigned}$$

$\rightarrow$ $165-2^8=-91_{two's\ complement}$

$\rightarrow$ $-2^8+1\times2^7+0\times2^6+1\times2^5+0\times2^4+0\times2^3+1\times2^2+0\times2^1+1\times2^0=-91$

$\rightarrow$ $-1\times2^7+0\times2^6+1\times2^5+0\times2^4+0\times2^3+1\times2^2+0\times2^1+1\times2^0=-91$

Signed Integers: Two's Complement

-2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
s	d_6	d_5	d_4	d_3	d_2	d_1	d_0

8 bit binary

$\rightarrow$ $$-s\times 2^7+\sum_{i=0}^{6} d_i 2^i$$

Range: $-128 \rightarrow +127$

$\pm$	2^6	2^5	2^4	2^3	2^2	2^1	2^0
1	0	1	0	0	1	0	1

$\rightarrow$ $$-1\times 2^7+0\times 2^6+1\times 2^5+0\times 2^4+0\times 2^3+1\times 2^2+0\times 2^1+1\times 2^0 =$$
$$= -91$$

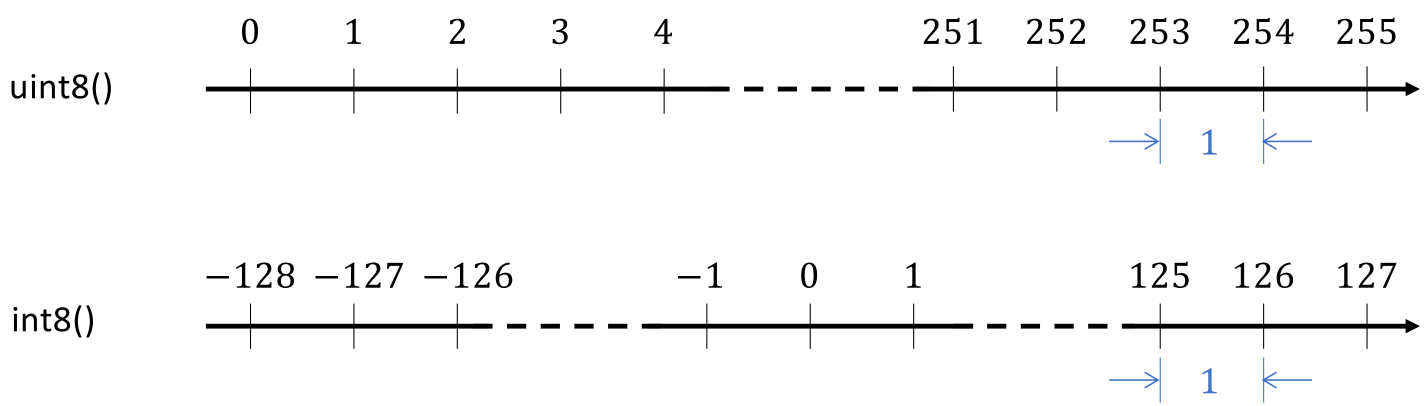
0	0	0	0	0	0	0	0

$\rightarrow$ 0

Integers

Type	Number of Bits	Range
uint8()	8	0 → 255
uint16()	16	0 → 65,535
uint32()	32	0 → 4,294,967,295
uint64()	64	0 → 18,446,744,073,709,551,615
int8()	8	−128 → 127
int16()	16	−32768 → 32767
int32()	32	−2,147,483,648 → 2,147,483,647
int64()	64	−9,223,372,036,854,775,808 → 9,223,372,036,854,775,807

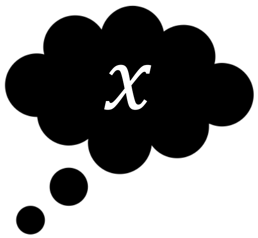
Set of Signed/Unsigned Integers



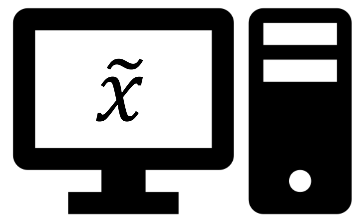
Distance between two adjacent stored values: 1

Number Representation

Abstract Quantity



Stored Quantity



$$x \in \mathbb{R}$$

$$0 \leq x \leq 4{,}294{,}967{,}295$$

$$\xrightarrow{\tilde{x} = uint32(x)}$$

$$\tilde{x} \in uint32$$

Error

Absolute Error: $\epsilon = |x - \tilde{x}|$

Relative Error: $\eta = \left|\frac{x - \tilde{x}}{x}\right|$

Rounding scheme → Round to nearest integer.

$$\epsilon = |x - \tilde{x}| \leq 0.5$$

What about relative error?

Fixed-Point Arithmetics

10^2	10^1	10^0		10^{-1}	10^{-2}	10^{-3}
1	3	2	.	0	4	5

$$\rightarrow \quad 1\times 10^2 + 3\times 10^1 + 2\times 10^0 + 0\times 10^{-1} + 4\times 10^{-2} + 5\times 10^{-3} =$$
$$= 132.045$$

Distance between two adjacent stored values : 10^{-3}

2^3	2^2	2^1	2^0		2^{-1}	2^{-2}	2^{-3}	2^{-4}
0	1	0	1	.	1	1	0	1

$$\rightarrow \quad 0\times 2^3 + 1\times 2^2 + 0\times 2^1 + 1\times 2^0 + \cdots$$
$$\cdots + 1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3} + 1\times 2^{-4} = 5.8125_{10}$$

Distance between two adjacent stored values : 2^{-4}

➔ Difficulty with dynamic range

Fixed-Point Arithmetics

- ✓ Can use similar (same) hardware as integer arithmetic.
- ✓ Fast computation.
- ✓ Numbers are equally spaced.

Numbers are equally spaced.

Loss of precision.

Limited Dynamic Range.

Floating Point Arithmetics

Example: Planck's Law

$$B_\nu(\nu,T)=\frac{2h\nu^3}{c^2}\frac{1}{e^{\frac{h\nu}{\kappa_B T}}-1}$$

Planck's Constant: $h=6.626070\times10^{-34}J.s$

Speed of light: $c=299{,}792{,}458m/s$

Boltzmann's constant: $\kappa_B=1.380649\times10^{-23}J/K$

For an IR wavelength of $1000nm$: $\nu\cong300THz=3\times10^{14}Hz$

Scientific Notation

Normalized value

$1123 \rightarrow 1.123\times10^{3}$

$11.23 \rightarrow 1.123\times10^{1}$

$0.1123 \rightarrow 1.123\times10^{-1}$

1.123×10^{30} $\quad$ 1.123×10^{-30}

Floating Point Example: Binary

Sign bit

Always ‘1’

Exponent

$$\pm 1.\underbrace{101001}_{\text{‘Significand’ or ‘Mantissa’}}\times 2^{3}$$

Floating Point: Binary

Sign bit
Always '1'
Exponent

$$\pm 1.M \times 2^{e}$$

'Significand' or 'Mantissa'

Float: Single Precision 32-bit

Sign bit

23-bit mantissa

b_{31}	b_{30}	b_{29}	b_{28}	b_{27}	b_{26}	b_{25}	b_{24}	b_{23}	b_{22}	b_{21}	b_{20}	b_{19}	b_{18}	b_{17}	b_{16}	b_{15}	b_{14}	b_{13}	b_{12}	b_{11}	b_{10}	b_9	b_8	b_7	b_6	b_5	b_4	b_3	b_2	b_1	b_0

8-bit exponent E

$$e = b_{30}b_{29}b_{28}b_{27}b_{26}b_{25}b_{24}b_{23} - \overbrace{127}^{\text{Bias}} = E - 127$$

Except for special cases (00000000 and 11111111)

➔ $-126 \leq e \leq 127$

Float: Single Precision 32-bit

Sign bit

23-bit mantissa M

b_{31}	b_{30}	b_{29}	b_{28}	b_{27}	b_{26}	b_{25}	b_{24}	b_{23}	b_{22}	b_{21}	b_{20}	b_{19}	b_{18}	b_{17}	b_{16}	b_{15}	b_{14}	b_{13}	b_{12}	b_{11}	b_{10}	b_9	b_8	b_7	b_6	b_5	b_4	b_3	b_2	b_1	b_0

8-bit exponent E

$$(-1)^{b_{31}}1.b_{22}b_{21}\ldots b_1 b_0 \times 2^{E-\overbrace{127}^{\text{Bias}}}$$

$$(-1)^{b_{31}}1.M \times 2^{E-127}$$

Special Cases, IEEE 754

$$b_{31}\underbrace{b_{30}\,b_{29}\,b_{28}\,b_{27}\,b_{26}\,b_{25}\,b_{24}\,b_{23}}_{E}\overbrace{b_{22}\,b_{21}\,b_{20}\,b_{19}\,b_{18}\,b_{17}\,b_{16}\,b_{15}\,b_{14}\,b_{13}\,b_{12}\,b_{11}\,b_{10}\,b_{9}\,b_{8}\,b_{7}\,b_{6}\,b_{5}\,b_{4}\,b_{3}\,b_{2}\,b_{1}\,b_{0}}^{M}$$

	$\boldsymbol{E=0}$	$\boldsymbol{0<E<255}$	$\boldsymbol{E=255}$
$\boldsymbol{M=0}$	± 0	$(-1)^{b_{31}}1.0\times 2^{E-127}$	$\pm\infty$
$\boldsymbol{M\neq 0}$	Denormalized	$(-1)^{b_{31}}1.M\times 2^{E-127}$	NaN

Distance between two adjacent numbers

$$\underbrace{+1.100101011100101010010}_{\text{23-bit mantissa}}\times 2^{0}$$

$$+1.100101011100101010011\times 2^{0}$$ Next higher number

Difference is $\epsilon_m = 2^{-23}\times 2^{0} \cong 1.1921\times 10^{-7}$

Valid for all x where $1 < x < 2$

Distance between two adjacent numbers

$$+1.\overbrace{10010101110010101010010}^{\text{23-bit mantissa}}\times 2^{1}$$

$$+1.10010101110010101010011\times 2^{1}$$

Next higher number

Difference is $\epsilon_m = 2^{-23}\times 2^{1} \cong 2.3842\times 10^{-7}$

Valid for all x where $2 < x < 4$

Distance between two adjacent numbers

$$+1.\overbrace{10010101110010101010010}^{\text{23-bit mantissa}}\times 2^{2}$$

$$+1.10010101110010101010011\times 2^{2}$$

Next higher number

Difference is $\epsilon_m = 2^{-23}\times 2^{2} \cong 4.7684\times 10^{-7}$

Valid for all x where $4 < x < 8$

Distance between two adjacent numbers

$$+1.\overbrace{10010101110010101010010}^{\text{23-bit mantissa}}\times 2^{127}$$

$$+1.10010101110010101010011\times 2^{127}$$

Next higher number

Difference is $|\Delta| = 2^{-23}\times 2^{127} \cong 2.0282\times 10^{31}$

Valid for all x where $2^{127} < x < 2^{128}$

Smallest Normalized Value

$$+1.\overbrace{00000000000000000000000}^{\text{23-bit mantissa}}\times 2^{-126}$$

$$+1.00000000000000000000001\times 2^{-126}$$

Next higher number

Difference is $\epsilon_m = 2^{-23}\times 2^{-126} \cong 1.1013\times 10^{-45}$

What about next smaller number?

Next number is zero if we do not de-normalize.

$\Rightarrow |\Delta| = 2^{-126} \cong 1.1755\times 10^{-38}$

Smallest Normalized Value

$$0$$

$$2^{-126} \cong 1.1755\times10^{-38}$$

$$+1.00000000000000000000000\times2^{-126}$$

$$2^{-23}\times2^{-126} \cong 1.4013\times10^{-45}$$

$$+1.00000000000000000000001\times2^{-126}$$

Denormalize $E = 0$

$$+0.11111111111111111111111\times 2^{-126}$$

$$2^{-23}\times 2^{-126} \cong 1.4013\times 10^{-45}$$

$$+1.00000000000000000000000\times 2^{-126}$$

$$2^{-23}\times 2^{-126} \cong 1.4013\times 10^{-45}$$

$$+1.00000000000000000000001\times 2^{-126}$$

Denormalized values $E = 0$

$$+0.11111111111111111111111 \times 2^{-126}$$

$$+0.00000000000000000000001 \times 2^{-126}$$

Half-Precision 16-bit

Sign bit

5-bit exponent $E = b_{14}b_{13}b_{12}b_{11}b_{10}$

10-bit mantissa $M = b_9 b_8 \ldots b_1 b_0$

b_{15}	b_{14}	b_{13}	b_{12}	b_{11}	b_{10}	b_9	b_8	b_7	b_6	b_5	b_4	b_3	b_2	b_1	b_0

$$e = b_{14}b_{13}b_{12}b_{11}b_{10} - 15 = E - \overbrace{15}^{\text{Bias}}$$

Except for special cases (00000 and 11111) ➔ $-14 \le e \le 15$

$$(-1)^{b_{15}} 1.M \times 2^{E-15}$$

Quarter Precision 8-bit (non-standard)

3-bit exponent $E = b_6 b_5 b_4$

Sign bit

4-bit mantissa $M = b_3 b_2 b_1 b_0$

b_7	b_6	b_5	b_4	b_3	b_2	b_1	b_0

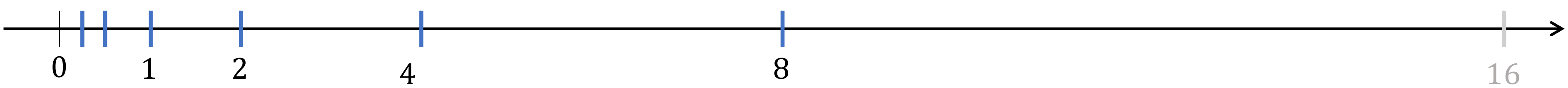
$$e = b_6 b_5 b_4 - 3 = E - \overbrace{3}^{\text{Bias}}$$

Except for special cases (000 and 111) ➔ $-2 \leq e \leq 3$

$$(-1)^{b_7} 1.M \times 2^{E-3}$$

Spacing (Powers of 2)

$M = 0000 \qquad (-1)^0 1.0000 \times 2^e \qquad -2 \leq e \leq 3$



$E = 110 \rightarrow e = 6 - 3 = 3 \rightarrow 2^3 = 8$

$E = 101 \rightarrow e = 5 - 3 = 2 \rightarrow 2^2 = 4$

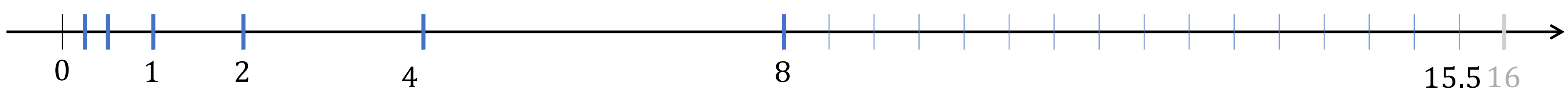
$\vdots$

$E = 010 \rightarrow e = 2 - 3 = -1 \rightarrow 2^{-1} = 0.5$

$E = 001 \rightarrow e = 1 - 3 = -2 \rightarrow 2^{-2} = 0.25$

Spacing ($E = 110$)

$(-1)^0 1.M \times 2^3 = 1b_3b_2b_1.b_0$ $\qquad 0000 \le M \le 1111$



$1000.0 \le \tilde{x} \le 1111.1$ $\qquad |\Delta| = 0.1_2 = 0.5_{10}$

$+1.0000 \times 2^3 = 1000.0_2 = 8_{10}$

$+1.0001 \times 2^3 = 1000.1_2 = 8.5_{10}$

$+1.0010 \times 2^3 = 1001.0_2 = 9_{10}$

$+1.0011 \times 2^3 = 1001.1_2 = 9.5_{10}$

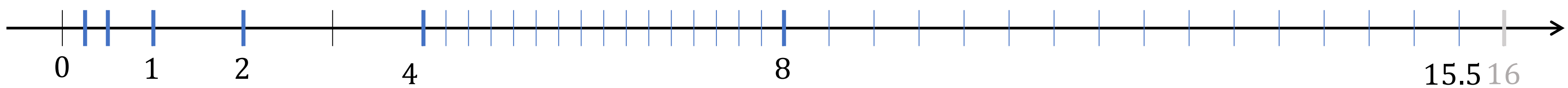
$+1.0100 \times 2^3 = 1010.0_2 = 10_{10}$

$\vdots$

$+1.1111 \times 2^3 = 1111.1_2 = 15.5_{10}$

Spacing ($E = 101$)

$(-1)^0 1.M \times 2^2 = 1b_3b_2.b_1b_0$ $\quad 0000 \leq M \leq 1111$



$100.00 \leq \tilde{x} \leq 111.11$ $\quad |\Delta| = 0.01_2 = 0.25_{10}$

$+1.0000\times 2^2 = 100.00_2 = 4_{10}$

$+1.0001\times 2^2 = 100.01_2 = 4.25_{10}$

$+1.0010\times 2^2 = 100.10_2 = 4.5_{10}$

$+1.0011\times 2^2 = 100.11_2 = 4.75_{10}$

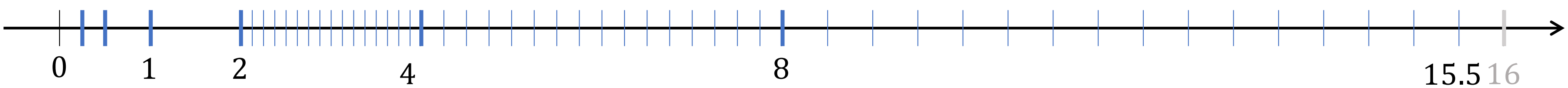
$+1.0100\times 2^2 = 101.00_2 = 5_{10}$

$\vdots$

$+1.1111\times 2^2 = 111.11_2 = 7.75_{10}$

Spacing ($E = 100$)

$$(-1)^0 1.M \times 2^1 = 1b_3.b_2\,b_1 b_0 \qquad 0000 \leq M \leq 1111$$



$$10.000 \leq \tilde{x} \leq 11.111 \qquad |\Delta| = 0.001_2 = 0.125_{10}$$

$$+1.0000 \times 2^1 = 10.000_2 = 2_{10}$$

$$+1.0001 \times 2^1 = 10.001_2 = 2.125_{10}$$

$$+1.0010 \times 2^1 = 10.010_2 = 2.25_{10}$$

$$+1.0011 \times 2^1 = 10.011_2 = 2.375_{10}$$

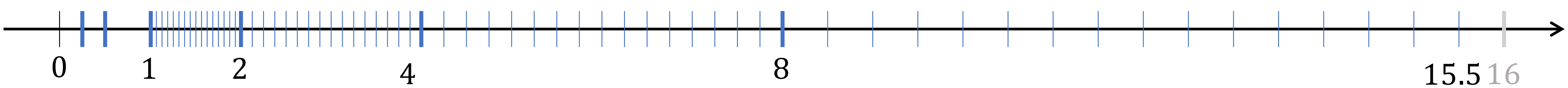
$$+1.0100 \times 2^1 = 10.100_2 = 2.5_{10}$$

$$\vdots$$

$$+1.1111 \times 2^2 = 11.111_2 = 3.875_{10}$$

Spacing ($E = 011$)

$(-1)^0 1.M \times 2^1 = 1.b_3 b_2 b_1 b_0$ $\quad$ $0000 \le M \le 1111$



$1.0000 \le \tilde{x} \le 1.1111$ $\quad$ $|\Delta| = 0.0001_2 = 2^{-4} = 0.0625_{10}$

$+1.0000 \times 2^0 = 1.0000_2 = 1_{10}$

$+1.0001 \times 2^0 = 1.0001_2 = 1.0625_{10}$

$+1.0010 \times 2^0 = 1.0010_2 = 1.125_{10}$

$+1.0011 \times 2^0 = 1.0011_2 = 1.1875_{10}$

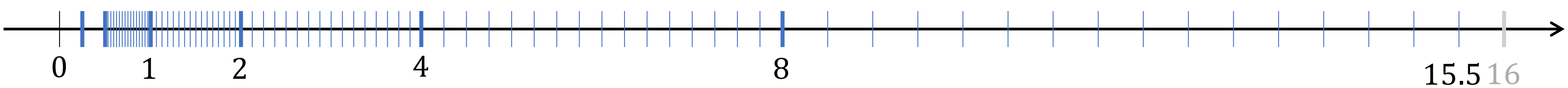
$+1.0100 \times 2^0 = 1.0100_2 = 1.25_{10}$

$\vdots$

$+1.1111 \times 2^0 = 1.1111_2 = 1.9375_{10}$

Spacing ($E = 010$)

$$(-1)^0 1.M \times 2^{-1} = 0.1b_3b_2b_1b_0 \qquad 0000 \le M \le 1111$$



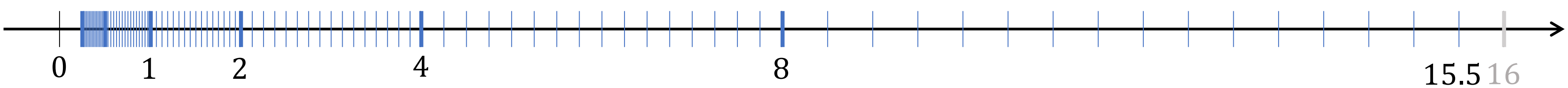
$$0.10000 \le \tilde{x} \le 0.11111 \qquad |\Delta| = 0.00001_2 = 2^{-5} = 0.03125_{10}$$

$$1.0000 \times 2^{-1} = 0.10000_2 = 0.5_{10}$$
$$1.0001 \times 2^{-1} = 0.10001_2 = 0.53125_{10}$$
$$1.0010 \times 2^{-1} = 0.10010_2 = 0.5625_{10}$$
$$1.0011 \times 2^{-1} = 0.10011_2 = 0.59375_{10}$$

$$1.0100 \times 2^{-1} = 0.10100_2 = 0.625_{10}$$
$$\vdots$$
$$1.1111 \times 2^{-1} = 0.11111_2 = 0.96875_{10}$$

Spacing ($E = 001$)

$(-1)^0 1.M \times 2^{-2} = 0.01 b_3 b_2 b_1 b_0 \quad 0000 \le M \le 1111$



$0.010000 \le \tilde{x} \le 0.011111 \qquad |\Delta| = 2^{-4} 2^{-2} = 2^{-6} = 0.015625_{10}$

$1.0000 \times 2^{-2} = 0.010000_2 = 0.25_{10}$

$1.0001 \times 2^{-2} = 0.010001_2 = 0.25 + 2^{-6}$

$1.0010 \times 2^{-2} = 0.010010_2 = 0.25 + 2 \times 2^{-6}$

$1.0011 \times 2^{-2} = 0.010011_2 = 0.25 + 3 \times 2^{-6}$

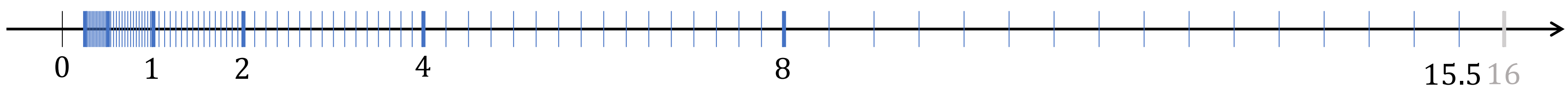
$0.10100 \times 2^{-2} = 0.25 + 4 \times 2^{-6}$

$\vdots$

$0.11111 \times 2^{-2} = 0.25 + 15 \times 2^{-6}$

Spacing ($E = 001$)

$$(-1)^0 1.M \times 2^{-2} \qquad 0000 \leq M \leq 1111$$



$$0.010000 \leq \tilde{x} \leq 0.011111 \qquad |\Delta| = 2^{-4}2^{-2} = 2^{-6} = 0.015625_{10}$$

$$1.0000 \times 2^{-2} = 0.010000_2 = 0.25_{10}$$

$$1.0001 \times 2^{-2} = 0.010001_2 = 0.25 + 2^{-6}$$

$$1.0010 \times 2^{-2} = 0.010010_2 = 0.25 + 2 \times 2^{-6}$$

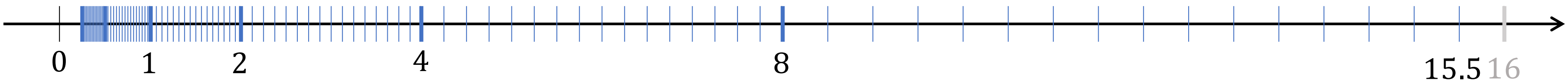
$$1.0011 \times 2^{-2} = 0.010011_2 = 0.25 + 3 \times 2^{-6}$$

$$0.10100 \times 2^{-2} = 0.25 + 4 \times 2^{-6}$$

$$\vdots$$

$$0.11111 \times 2^{-2} = 0.25 + 15 \times 2^{-6}$$

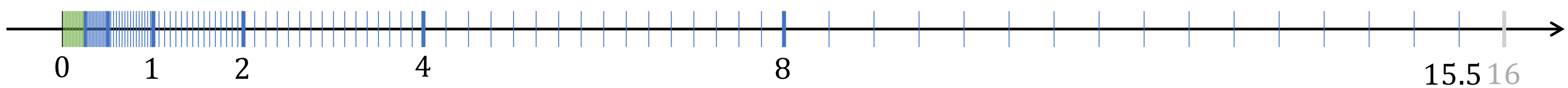
Normalized Positive Values



- Smallest normalized value: 0.25
- Largest normalized value: 15.5
- Spacing is the same between powers of 2
- There are 16 equally spaced values at each value of E
- There are $6 \times 16 = 96$ normalized positive floating-point values
- There is a 'large' gap between zero and the smallest normalized value

Subnormal Positive Values ($E = 000$)

$$(-1)^0 0.M\times 2^{-2} = 0.00b_3b_2b_1b_0 \quad 0001 \leq M \leq 1111$$

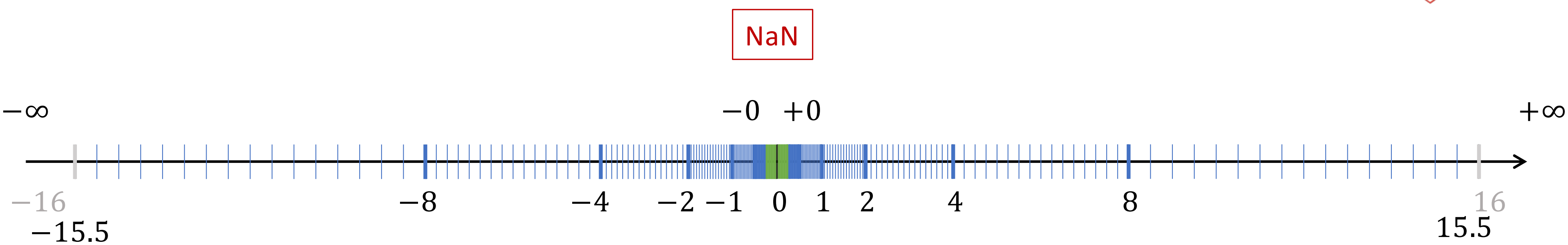


$$0.000001 \leq \tilde{x} \leq 0.001111 \qquad |\Delta| = 2^{-4}2^{-2} = 2^{-6} = 0.015625_{10}$$

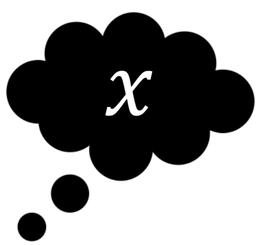
Special Cases, “8-bit Floating Point”

	$\boldsymbol{E=0}$	$\boldsymbol{0<E<7}$	$\boldsymbol{E=7}$
$\boldsymbol{M=0}$	± 0	$(-1)^{b_7}1.0\times 2^{E-3}$	$\pm\infty$
$\boldsymbol{M\neq 0}$	Subnormal	$(-1)^{b_7}1.M\times 2^{E-3}$	NaN

Set of Floating-Point Values $\mathbb{F}$

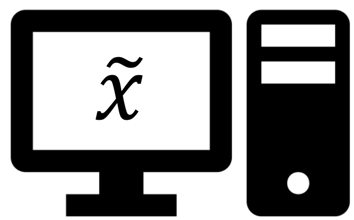


Abstract Quantity



$x \in \mathbb{R}$

Stored Quantity



$\tilde{x} \in \mathbb{F}$

Floating Point

Type	Bits	Exponent	Mantissa	Exponent Bias	e_{min}	e_{max}
half()	16	5 bits	10 bits	15	-14	15
single()	32	8 bits	23 bits	127	-126	127
double()	64	11 bits	52 bits	1,023	-1,022	1,023

Double() Largest Value

$$\mathrm{M} = \underbrace{111\cdots11}_{\text{52 bits}} \qquad E = \overbrace{111\cdots10}^{\text{11 bits}} \qquad e = 2046 - 1023 = 1023$$

$$+1.111\cdots11\times2^{1023} = \quad +\underbrace{1111\cdots11}_{\text{53 bits}}\times2^{971} =$$

$$= 1.797693134862316\times10^{308}$$

Double() Smallest Normal Value

$$\mathrm{M} = \underbrace{000\cdots00}_{52\text{ bits}} \qquad E = \overbrace{000\cdots01}^{11\text{ bits}} \qquad e = 1 - 1023 = -1022$$

$$+1.000\cdots00\times2^{-1022} = 2.225073858507201\times10^{-308}$$

Double() Smallest Subnormal Value

$$\mathrm{M} = \underbrace{000\cdots01}_{\text{52 bits}} \qquad E = \overbrace{000\cdots00}^{\text{11 bits}} \qquad e = -1022$$

$$+0.000\cdots01\times2^{-1022} = 2^{-1074}$$

$$= 4.940656458412465\times10^{-324}$$