Floating Point Number

Floating Point Representation

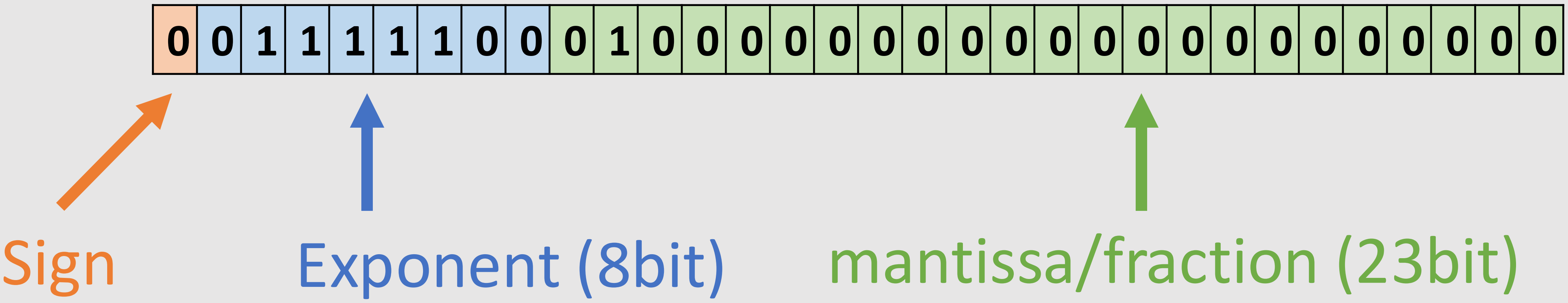
$$
\begin{aligned}
(5.625)_{10} &= 4 + 1 + 0.5 + \frac{1}{8} \\
&= 1 \cdot 2^{2} + 0 \cdot 2^{1} + 1 \cdot 2^{0} + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} \ldots \\
&= (101.101)_{2} \\
&= (1.01101)_{2} \cdot 2^{2}
\end{aligned}
$$

exponent

radix

mantissa/fraction

Floating Point Expression in IEEE754 (f32)



0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

=0.15625

Sign

Exponent (8bit)

mantissa/fraction (23bit)

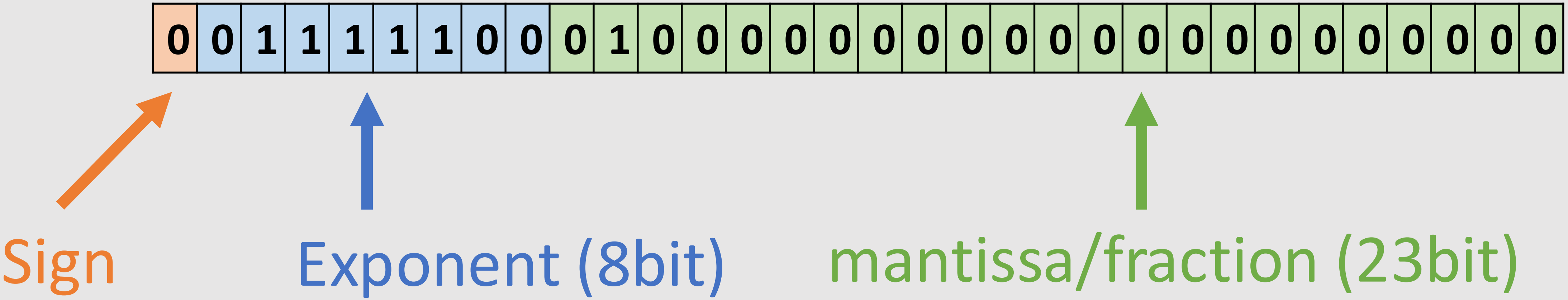
$$\boxed{Value = (-1)^{sign} \cdot 2^{exponent-127} \cdot (1.fraction)}$$

$1\ or\ -1$

$[2^{-127}, 2^{128}]$

$[1,2)$

Floating Point Expression in IEEE754 (f32)



0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

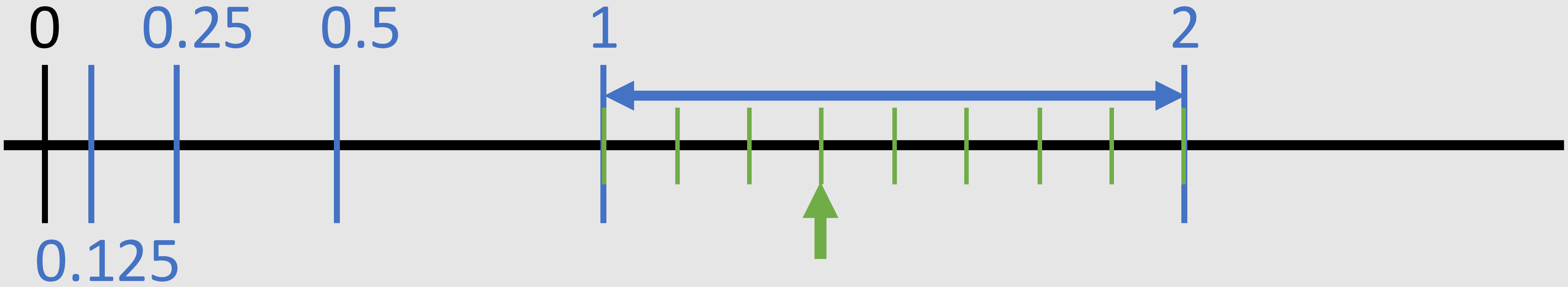
=0.15625

Sign

Exponent (8bit)

mantissa/fraction (23bit)

$$\boxed{Value = (-1)^{sign} \cdot 2^{exponent-127} \cdot (1.fraction)}$$



Special Value

$$\boxed{Value = (-1)^{sign} \cdot 2^{exponent-127} \cdot (1.fraction)}$$

Zero	?	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Infinity	?	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NaN	?	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Representing Integer in Floating Point

$$
\begin{aligned}
(27)_{10} &= 16 + 8 + 2 + 1 \\
&= 1 \cdot 2^{4} + 1 \cdot 2^{3} + 0 \cdot 2^{2} + 1 \cdot 2^{1} + 1 \cdot 2^{0} + 0 \cdot 2^{-1} \ldots \\
&= (11011.00000)_{2} \\
&= (1.10110000000)_{2} \cdot 2^{4}
\end{aligned}
$$

exponent

radix

mantissa/fraction

A floating point is integer when the mantissa is all zero after N-th most significant bit, where exponent is N