Floating point number

Datatype	FP64	FP32	TF32	FP16	BF16	FP8e5m2	FP8e4m3
Sign bit	1	1	1	1	1	1	1
Exponent (k)	11	8	8	5	8	5	4
Mantissa (n)	52	23	10	10	7	2	3

- Maximum (minimum)

Datatype	FP64	FP32	TF32	FP16	BF16	FP8e5m2	FP8e4m3
Sign bit ($s_{\bar{2}}$)	0	0	0	0	0	0	0
Exponent ($e_{\bar{2}}$)	11111111110	11111110	11111110	11110	11111110	11110	1110
Mantissa ($m_{\bar{2}}$)	111...1	111...1	111...1	111...1	1111111	11	111
$E = 2^e - \left(2^{k-1} - 1\right)$	1023	127	127	15	127	15	7
$M = \sum_{i=1}^{n} \frac{1}{2^i}$	$1 - \frac{1}{2^{52}}$	$1 - \frac{1}{2^{23}}$	$1 - \frac{1}{2^{10}}$	$1 - \frac{1}{2^{10}}$	$1 - \frac{1}{2^7}$	$1 - \frac{1}{2^2}$	$1 - \frac{1}{2^3}$
$max = (-1)^s 2^E \left(1 + M\right)$	1.798×10^{308}	3.403×10^{38}	3.401×10^{38}	$65504.$	3.390×10^{38}	$57344.$	$240.$
Sign bit of minimum ($s_{\bar{2}}$)	1	1	1	1	1	1	1
$min = (-1)^s 2^E \left(1 + M\right)$	-1.798×10^{308}	-3.403×10^{38}	-3.401×10^{38}	$-65504.$	-3.390×10^{38}	$-57344.$	$-240.$

- Absolute minimum

Datatype	FP64	FP32	TF32	FP16	BF16	FP8e5m2	FP8e4m3
Sign bit ($s_{\bar{2}}$)	0	0	0	0	0	0	0
Exponent ($e_{\bar{2}}$)	00000000000	00000000	00000000	00000	00000000	00000	0000
Mantissa ($m_{\bar{2}}$)	000...01	000...01	000...01	000...01	0000001	01	001
$E = 1 - \left(2^{k-1} - 1\right)$	-1022	-126	-126	-14	-126	-14	-6
$M = \frac{1}{2^n}$	$\frac{1}{2^{52}}$	$\frac{1}{2^{23}}$	$\frac{1}{2^{10}}$	$\frac{1}{2^{10}}$	$\frac{1}{2^7}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$
$value = (-1)^s 2^E M$	4.941×10^{-324}	1.401×10^{-45}	1.148×10^{-41}	5.960×10^{-8}	9.184×10^{-41}	1.526×10^{-5}	1.953×10^{-3}

- Other value

Datatype	$+\infty$	$-\infty$	NaN
Sign bit ($s_{\bar{2}}$)	0	1	0 or 1
Exponent ($e_{\bar{2}}$)	all 1	all 1	all 0
Mantissa ($m_{\bar{2}}$)	all 0	all 0	not all 0

Signed integer

Datatype	INT64	INT32	INT8	INT4
Maximum	$2^{63} - 1$	$2^{31} - 1$	$2^7 - 1$	$2^3 - 1$
value of maximum	$9223372036854775807 \approx 9.2 \times 10^{18}$	$2147483647 \approx 2.1 \times 10^9$	32767	7