

Dr. Martin Köhler, Dr. Pavan Laxmipathy Veluvali

Website: <http://www.mpi-magdeburg.mpg.de/scicomp2425>

Scientific Computing 1 Handout 6 October 16, 2024

The IEEE 754 Standard

precision	p	t	e_{\min}	e_{\max}	u	x_{\min}	x_{\max}
half	2	$10 + 1$	-13	16	$\approx 4.88 \cdot 10^{-4}$	$\approx 6 \cdot 10^{-5}$	$\approx 1 \cdot 10^5$
single	2	$23 + 1$	-125	128	$\approx 5.96 \cdot 10^{-8}$	$\approx 1 \cdot 10^{-38}$	$\approx 3 \cdot 10^{38}$
double	2	$52 + 1$	-1021	1024	$\approx 1.11 \cdot 10^{-16}$	$\approx 10^{-308}$	$\approx 10^{308}$
quad	2	$112 + 1$	-16381	16384	$\approx 9.63 \cdot 10^{-35}$	$\approx 10^{-4932}$	$\approx 10^{4932}$

Table 1: IEEE standard 754-2008, data types.

Storage patterns for half, single and double precision variables:

half: (16 bit)

```
S EEEEE MMMMMMMMM
0 1   5 6       15
```

single: (32 bit)

```
S EEEEEEEE MMMMMMMMMMMMMMMMMMMMMMMMMMMMM
0 1         8 9                               31
```

double: (64 bit)

```
S EEEEEEEEEEE MMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM
0 1           11 12                                                                           63
```

Flag	Example	Result
<i>invalid</i>	$0/0, 0 \cdot \infty, \sqrt{-1}, \infty/\infty, +\infty + (-\infty)$	NaN ("not a number")
<i>overflow</i>	$x_{\max} * x_{\max}$	$\pm\infty$ usually denoted: $\pm\text{Inf}$
<i>division by zero</i>	$x/0$ for $x \neq 0$	$\pm\infty$
<i>underflow</i>	$x_{\min}/p^s, 1 < s < t$	subnormal number
<i>inexact</i>	$\text{rd}(x \circ y) \neq x \circ y$	correctly rounded result

Table 2: IEEE Standard 754, Exception Handling.

Examples:

0	11111111	000000000000000000000000	=	$+\infty$
1	11111111	000000000000000000000000	=	$-\infty$
<hr/>				
0	11111111	000001000000000000000000	=	NaN
1	11111111	00100010001001010101010	=	NaN
<hr/>				
0	10000000	000000000000000000000000	=	$+1.0 * 2^{128-127} = 2$
0	10000001	101000000000000000000000	=	$+1.101 * 2^{129-127} = 6.5$
1	10000001	101000000000000000000000	=	$-1.101 * 2^{129-127} = -6.5$
<hr/>				
0	00000001	000000000000000000000000	=	$+1.0 * 2^{1-127} = 2^{-126} = x_{\min}$
0	00000000	100000000000000000000000	=	$+0.1 * 2^{-126} = 2^{-127}$
0	00000000	000000000000000000000001	=	$+0.0\dots01 * 2^{-126} = 2^{-149}$ = smallest representable number
<hr/>				
0	00000000	000000000000000000000000	=	+0
1	00000000	000000000000000000000000	=	-0
<hr/>				
0	01111111	000000000000000000000000	=	$+1.0 * 2^{127-127} = +1.0$
1	01111111	000000000000000000000000	=	$-1.0 * 2^{127-127} = -1.0$