~~0100110000001010~~

① $\underbrace{0}_{\text{sign}}\,|\,\underbrace{10011}_{\text{exponent}}\,|\,\underbrace{0000001010}_{\text{significand}}$

~~0100110000001011~~

Exponent: $2^0 \times 1 + 2^1 \times 1 + 2^2 \times 0 + 2^3 \times 0 + 2^4 \times 1$

$= 1 + 2 + 16$

$= 19.$

subtract bias

Power of 2: $19 - 15 = 4$

Significand

$1.0000001010 \times 2^{19-15}$

$2^{-1} \times 0 \quad 2^{-2} \times 0 \quad 2^{-7} \times 1 \quad 2^{-9} \times 1$

Fraction

$$\left(1 + \frac{1}{2^7} + \frac{1}{2^9}\right) \times 2^4$$

$$= \left(1 + \frac{1}{128} + \frac{1}{512}\right) \times 2^4$$

$$= \left(\frac{512}{512} + \frac{4}{512} + \frac{1}{512}\right) \times 2^4$$

$$= \frac{517}{512} \times 16 = \boxed{\frac{517}{32}} \approx \boxed{16.16}$$

② The next biggest number should increment the significand, leaving the same exponent.

a) 0|10011|0000001011

Here, we used the fact that 1011 is the next biggest binary number after 1010

(b) and (c) using a similar procedure:

$$1.0000001011 \times 2^{19-15}$$

$2^{-1} \times 0 \quad 2^{-2} \times 0 \quad 2^{-7} \times 1 \quad 2^{-9} \times 1 \quad 2^{-10} \times 1$

$$= \left(1 + \frac{1}{2^7} + \frac{1}{2^9} \times \frac{1}{2^{10}}\right) \times 2^4$$

$$= \left(\frac{1024}{1024} + \frac{8}{1024} + \frac{2}{1024} + \frac{1}{1024}\right) \times 2^4$$

$$= \frac{1035}{1024} \times 2^4 = \boxed{\frac{1035}{64}} \approx \boxed{16.17}$$

Machine Epsilon : difference between 1 and next bigge number.

0011110000000000
0011110000000001 $\longrightarrow (1+2^{-10}) - 1 = 1/2^{10} \approx 9.7\times10^{-4}$

00111111100...0
$\underbrace{00111111100...1}_{\text{23 bits}} \longrightarrow (1+2^{-23}) - 1 = 1/2^{23} \approx 1.2\times10^{-7}$

0011111111100...0
$\underbrace{0011111111100...1}_{\text{52 bits}} \longrightarrow (1+2^{-52}) - 1 = 1/2^{52} \approx 2.2\times10^{-16}$

Largest number

We set both exponent and significand to be the highest they can be without showing up as "infinity" or "NaN".

~~0111~~1011111111111

$= 1.111....1 \times 2^{15}$

$= \left(1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\dots+\frac{1}{1024}\right)\times 2^{15} = \frac{2047}{1024}\times 2^{15}$

$= 65{,}504$

Exponent = 30
minus bias = 15

$\underbrace{011111110}\underbrace{11...1}_{\text{23 bits}}$

$= 1.111...1 \times 2^{127}$

$= \left(1+\frac{1}{2}+\frac{1}{4}+\dots+\frac{1}{2^{23}}\right)\times 2^{127}$

Exponent = 254
minus bias = 127

$= 1.999.... \times 2^{15}$

$\approx 2\times2^{127} = 2^{128}$

$= 3.4\times10^{38}$

$\underbrace{011111111110}\underbrace{11..1}_{\text{52 bits}}$

$= 1.11...1 \times 2^{5}$

$= \left(1+\frac{1}{2}+\frac{1}{4}+\dots+\frac{1}{2^{52}}\right)\times 2^{1023}$

Exponent = 2046
minus bias = 1023

$= 1.99... \times 2^{1023}$

$\approx 2\times2^{1023} = 2^{1024}$

$= 1.8\times10^{308}$

Smallest Number (greater than zero)

$000\ldots1$

significand : $00\cdots1$
exponent : $00\cdots0$

General form

$$\boxed{0.00\ldots1 \times 2^{0-\text{bias}+1}}$$

16-bit $0.\overbrace{00\ldots1}^{10} \times 2^{-15+1}$

$$= \left(0+\frac{1}{2^{10}}\right) \times 2^{-14} = 2^{-24} \approx 5.97 \times 10^{-8}$$

32-bit $0.\overbrace{00\ldots.1}^{23} \times 2^{-127+1}$

$$= \left(0+\frac{1}{2^{23}}\right) \times 2^{-126} = 2^{-149} \approx 1.4 \times 10^{-45}$$

64-bit $0.\overbrace{00\ldots1}^{52} \times 2^{-1023+1}$

$$= \left(0+\frac{1}{2^{52}}\right) \times 2^{-1022} = 2^{-1074} \approx 5 \times 10^{-324}$$

Number of Numbers

including zero but excluding ∞ and NaN.

16-bit : possible numbers: 2^{16}

When exponent is all 1's, we don't have numbers

$$\underset{1\text{ bit}}{0}\,11111\,\underbrace{0000000000}_{10\text{ bits}}$$

Whether these are zero or one, no admissable numbers. 2^{11} possible numbers here.

$$\Rightarrow 2^{16} - 2^{11} = 63{,}488$$

32-bit possible numbers: 2^{32}

When exponent is all 1's, we don't have numbers

$$\underset{1\text{ bit}}{0}\,11111\,\underbrace{00000\cdots000}_{23\text{ bits}}$$

Whether these are zero or one, no admissable numbers. 2^{24} possible numbers here.

$$\Rightarrow 2^{32} - 2^{24} \approx 4.27\times10^{9}$$

64-bit possible numbers: 2^{64}

When exponent is all 1's, we don't have numbers

$$\underset{1\text{ bit}}{0}\,11111\,\underbrace{00000\cdots000}_{52\text{ bits}}$$

Whether these are zero or one, no admissable numbers. 2^{53} possible numbers here.

$$\Rightarrow 2^{64} - 2^{53} \approx 1.84\times10^{19}$$

How many 16-bit floats between 2^4 and 2^5?

$2^4 = 16$:

19

bias : 15

$1.000\ldots0_2 \times 2^{19-15}$

The next number will increment significand.

*

**

$1.111\ldots1_2 \times 2^4$

and so on. There are 10 binary digits in significand.

$\Rightarrow$ 2^{10} numbers b/w 2^4 and 2^5.

Gap Size

Difference b/w * and ** is :

$$\overbrace{\left(1 + \frac{1}{2^9}\right) \times 2^4}^{*} - \overbrace{\left(1 + \frac{1}{2^{10}}\right) \times 2^4}^{**}$$

can cancel 1's

$$= 2^4 \left[\frac{1}{2^9} - \frac{1}{2^{10}}\right]$$

$$= 2^4 \left[\frac{2}{2^{10}} - \frac{1}{2^{10}}\right] = 2^4 \times \frac{1}{2^{10}} = \boxed{2^{-6}} \approx 0.0156$$

How many 16-bit floats between 2^7 and 2^8?

$\underbrace{2^7}_{22} = \underbrace{128:}$

bias: 15

$1.000\ldots0_2 \times 2^{22-15}$

The next number will increment significand.

$\ast$

$\ast\ast$

$1.111\ldots1_2 \times 2^4$

and so on. There are 10 binary digits in significand.
$\Rightarrow$ 2^{10} numbers b/w 2^7 and 2^8.

Gap Size

Difference b/w $\ast$ and $\ast\ast$ is:

$$\overbrace{\left(1 + \frac{1}{2^9}\right) \times 2^7}^{\ast} - \overbrace{\left(1 + \frac{1}{2^{10}}\right) \times 2^7}^{\ast\ast}$$ can cancel 1's

$$= 2^7 \left[\frac{1}{2^9} - \frac{1}{2^{10}}\right]$$

$$= 2^7 \left[\frac{2}{2^{10}} - \frac{1}{2^{10}}\right] = 2^7 \times \frac{1}{2^{10}} = \boxed{2^{-3}} = 0.125$$

How many 16-bit floats between 2^{-5} and 2^{-4} ?

$$2^{-5} = 1/32$$

$$1.000\ldots00_2 \times 2^{-5}$$

10 . minus bias = -5

The next number will increment significand.

$\ast$

$\ast\ast$

$\rightarrow$ biggest number with exponent -5

$\Rightarrow$ 2^{10} 16-bit floats between 2^{-5} and 2^{-4}.

Gap Size

Difference b/w $\ast$ and $\ast\ast$ is :

$$\overbrace{\left(1 + \frac{1}{2^9}\right) \times 2^{-5}}^{\ast} - \overbrace{\left(1 + \frac{1}{2^{10}}\right) \times 2^{-5}}^{\ast\ast}$$

can cancel 1's

$$= 2^{-5}\left[\frac{1}{2^9} - \frac{1}{2^{10}}\right]$$

$$= 2^{-5}\left[\frac{2}{2^{10}} - \frac{1}{2^{10}}\right] = 2^{-5} \times \frac{1}{2^{10}} = \boxed{2^{-15}} \approx 0.0000305$$

How many 16-bit floats between 2^{-8} and 2^{-7} ?

$2^{-8} = 1/256$

$1.000\ldots00_2 \times 2^{-8}$

7. minus bias = −8

The next number will increment significand.

★

★★

→ biggest number with exponent −8

$\Rightarrow 2^{10}$ 16-bit floats between 2^{-8} and 2^{-7}.

Gap Size

Difference b/w ★ and ★★ is :

$$\overbrace{\left(1 + \frac{1}{2^9}\right) \times 2^{-8}}^{★} - \overbrace{\left(1 + \frac{1}{2^{10}}\right) \times 2^{-8}}^{★★}$$

can cancel 1's

$$= 2^{-8}\left[\frac{1}{2^9} - \frac{1}{2^{10}}\right]$$

$$= 2^{-8}\left[\frac{2}{2^{10}} - \frac{1}{2^{10}}\right] = 2^{-8} \times \frac{1}{2^{10}} = \boxed{2^{-18}} \approx 0.00000381$$