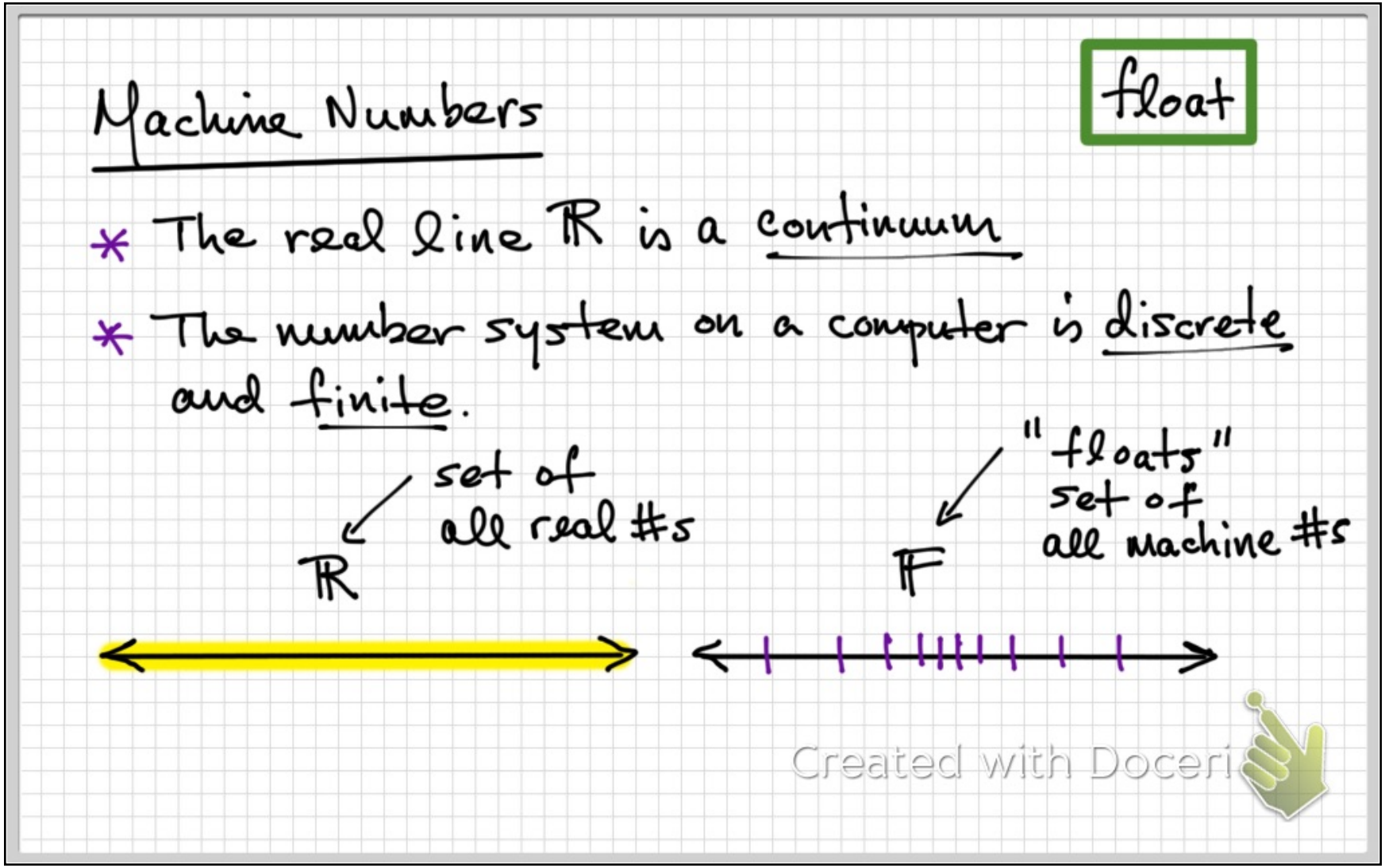

Machine Numbers
float
* The real line ℝ is a continuum
* The number system on a computer is discrete and finite.
ℝ
set of all real #s
𝔽
"floats" set of all machine #s
Created with Doceri

Mostly, given $x \in \mathbb{R}$ of interest, there is an $fl(x) \in \mathbb{F}$ which represents x on the computer. How do we find $fl(x)$?

$fl(x)$ is closely related to binary representation of x.

Problem Find $fl(9.4)$

→ Step ① : find binary rep of 9.4

focus on this step now

Step ② : express in scientific notation

step ③ : truncate/round if necessary

Recall: $x = (b_M \cdots b_2 b_1 b_0 . b_{-1} b_{-2} \cdots\cdots)_2$

$= b_M 2^M + \cdots + b_2 2^2 + b_1 2^1 + b_0 2^0$

$+ b_{-1} 2^{-1} + b_{-2} 2^{-2} + \cdots$

Where each integer bit is either 0 or 1.

$9.4 = 9 + 0.4$

integer part

fractional part

Convert these to binary separately.

Integer part $9 = b_m 2^m + \cdots + b_1 2^1 + b_0 2^0$

$9/2 = 4$ remainder 1 (get b_0)

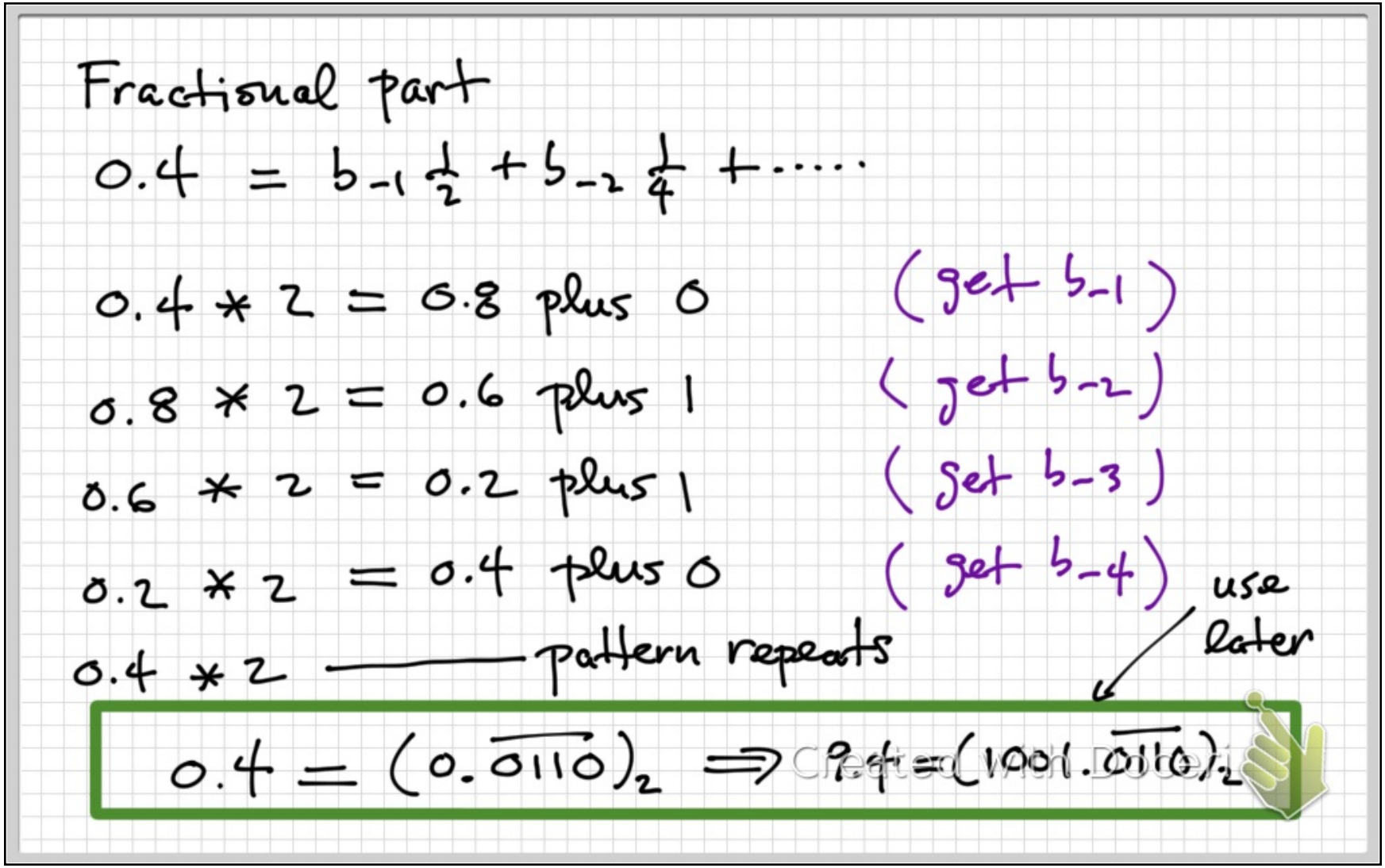
$4/2 = 2$ remainder 0 (get b_1)

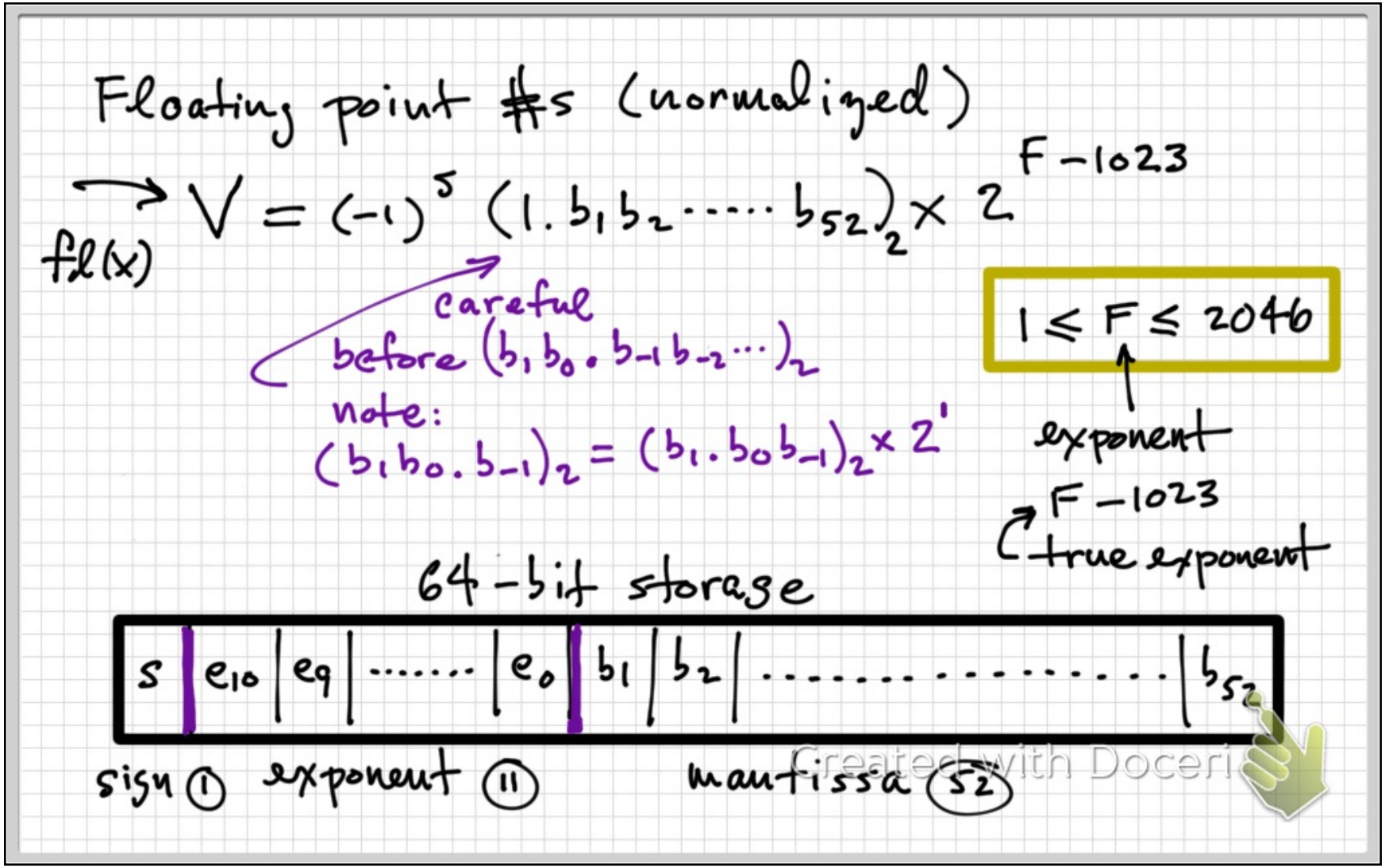
$2/2 = 1$ remainder 0 (get b_2)

$1/2 = 0$ remainder 1 (get b_3 and stop)

$9 = (1001)_2$, $1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$

$= 8 + 0 + 0 + 1 = 9$


Fractional part
$0.4 = b_{-1}\frac{1}{2} + b_{-2}\frac{1}{4} + \cdots$
$0.4 * 2 = 0.8$ plus 0 (get b_{-1})
$0.8 * 2 = 0.6$ plus 1 (get b_{-2})
$0.6 * 2 = 0.2$ plus 1 (get b_{-3})
$0.2 * 2 = 0.4$ plus 0 (get b_{-4})
$0.4 * 2$ —— pattern repeats
use later
$0.4 = (0.\overline{0110})_2 \Rightarrow$ Ref $= (1001.\overline{0110})_2$


Floating point #s (normalized)
fl(x)
$V = (-1)^s (1.b_1 b_2 \cdots b_{52})_2 \times 2^{F-1023}$
careful
before $(b_1 b_0 . b_{-1} b_{-2} \cdots)_2$
note:
$(b_1 b_0 . b_{-1})_2 = (b_1 . b_0 b_{-1})_2 \times 2^1$
$1 \leq F \leq 2046$
exponent
$F-1023$
true exponent
64-bit storage
s
e_{10}
e_9
e_0
b_1
b_2
b_{52}
sign (1)
exponent (11)
mantissa (52)
Created with Doceri

$$V = (-1)^S (1.b_1 b_2 \cdots\cdots b_{52})_2 \times 2^{F-1023}$$

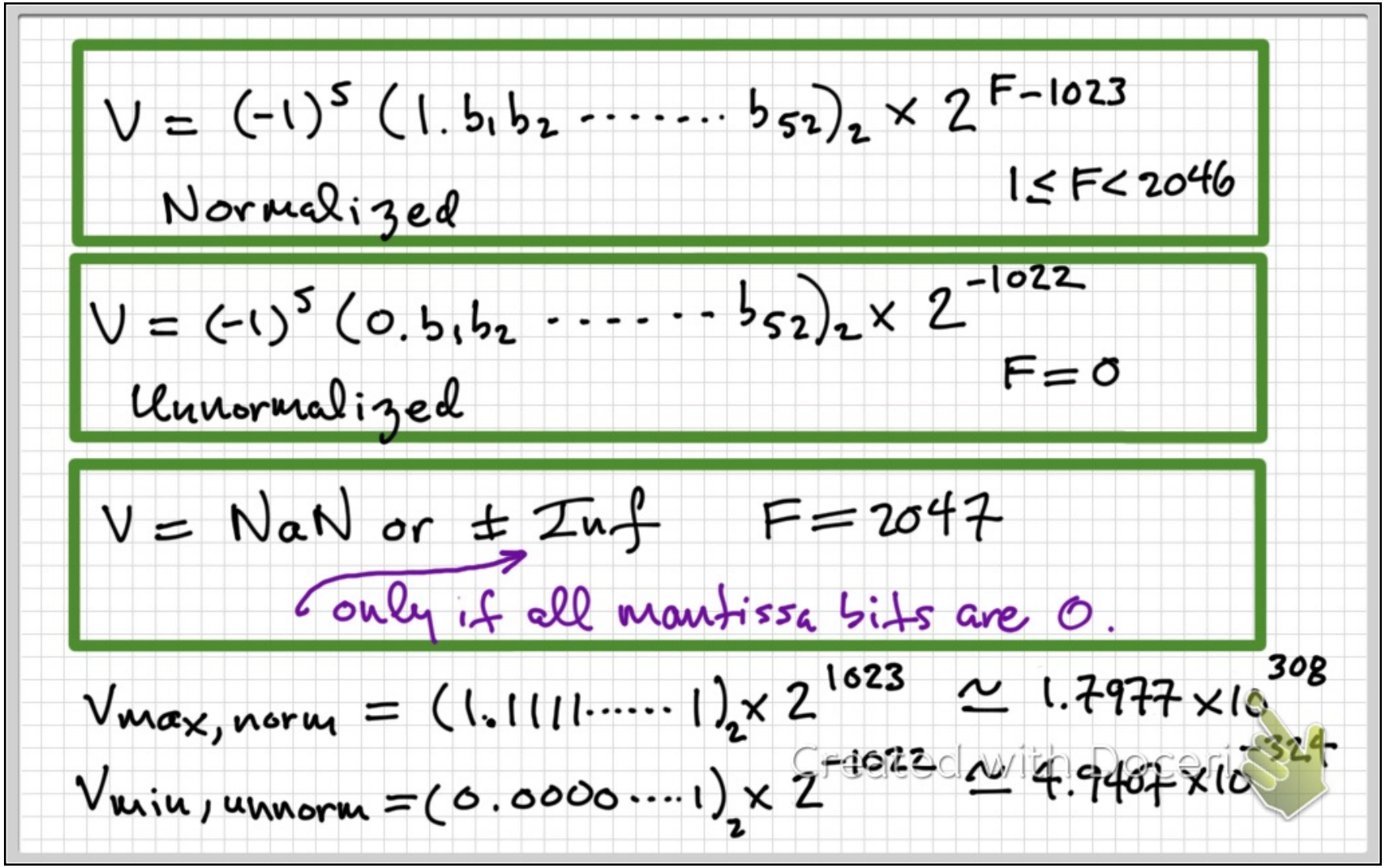
$1 \le F \le 2046$ normalized #s

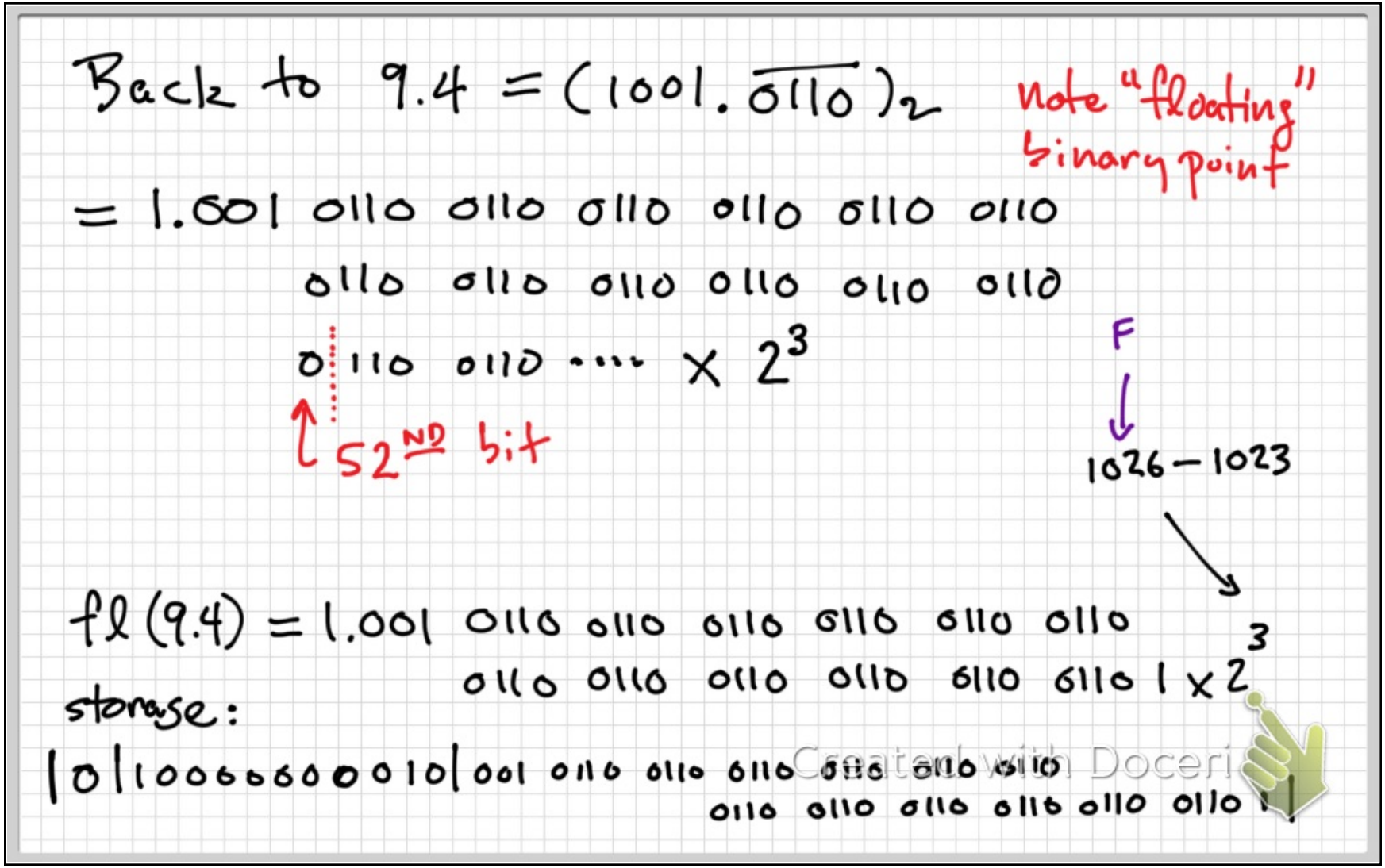
$F = 0$ unnormalized #s

$F = 2047$ NaNs and $\pm$ Inf

$F = (e_{10} e_9 e_8 e_7 e_6 e_5 e_4 e_3 e_2 e_1 e_0)_2$

note $= (\underbrace{11111111111}_{11 \text{ ones}})_2 = 2^{11} - 1 = 2048 - 1 = 2047$ ✓


$V = (-1)^S (1.b_1 b_2 \cdots\cdots b_{52})_2 \times 2^{F-1023}$
Normalized
$1 \le F < 2046$
$V = (-1)^S (0.b_1 b_2 \cdots\cdots b_{52})_2 \times 2^{-1022}$
Unnormalized
$F = 0$
$V = NaN$ or $\pm Inf$ $F = 2047$
only if all mantissa bits are 0.
$V_{max,norm} = (1.1111\cdots\cdots 1)_2 \times 2^{1023} \simeq 1.7977 \times 10^{308}$
$V_{min,unnorm} = (0.0000\cdots 1)_2 \times 2^{-1022} \simeq 4.9407 \times 10^{-324}$
Created with Doceri


Back to 9.4 = (1001.0110)2
note "floating" binary point
= 1.001 0110 0110 0110 0110 0110 0110
0110 0110 0110 0110 0110 0110
0110 0110 .... × 2^3
52ND bit
F
1026 − 1023
fl(9.4) = 1.001 0110 0110 0110 0110 0110 0110
0110 0110 0110 0110 0110 0110 1 × 2^3
storage:
0 10000000010 001 0110 0110 0110 0110 0110 0110
0110 0110 0110 0110 0110 0110
Created with Doceri

Notice

|0|10000000010|001 0110 0110 0110 0110 0110 0110 0110 0110 0110 0110 0110 1|

is

|0100 0000 0010 0010 1100 1100 1100 1100 1100 1100 1100 1100 1100 1100 1101|

4022cccccccccccd

$(0)_{16} = (0000)_2$ $(4)_{16} = (0100)_2$ $(8)_{16} = (1000)_2$ $(c)_{16} = (1100)_2$

$(1)_{16} = (0001)_2$ $(5)_{16} = (0101)_2$ $(9)_{16} = (1001)_2$ $(d)_{16} = (1101)_2$

$(2)_{16} = (0010)_2$ $(6)_{16} = (0110)_2$ $(a)_{16} = (1010)_2$ $(e)_{16} = (1110)_2$

$(3)_{16} = (0011)_2$ $(7)_{16} = (0111)_2$ $(b)_{16} = (1011)_2$ $(f)_{16} = (1111)_2$