More on Machine Numbers

float

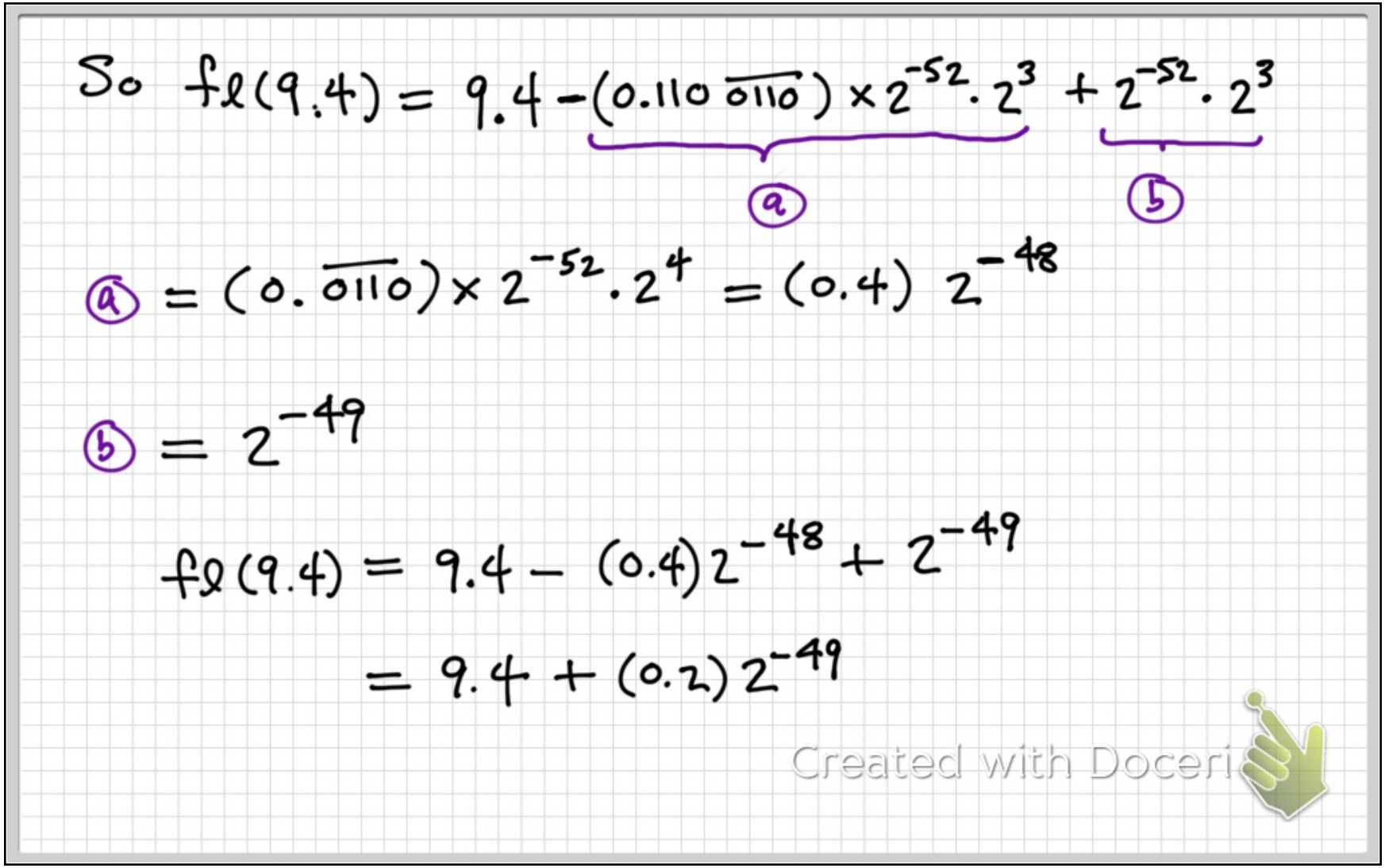
$9.4 = (1001.\overline{0110})_2$

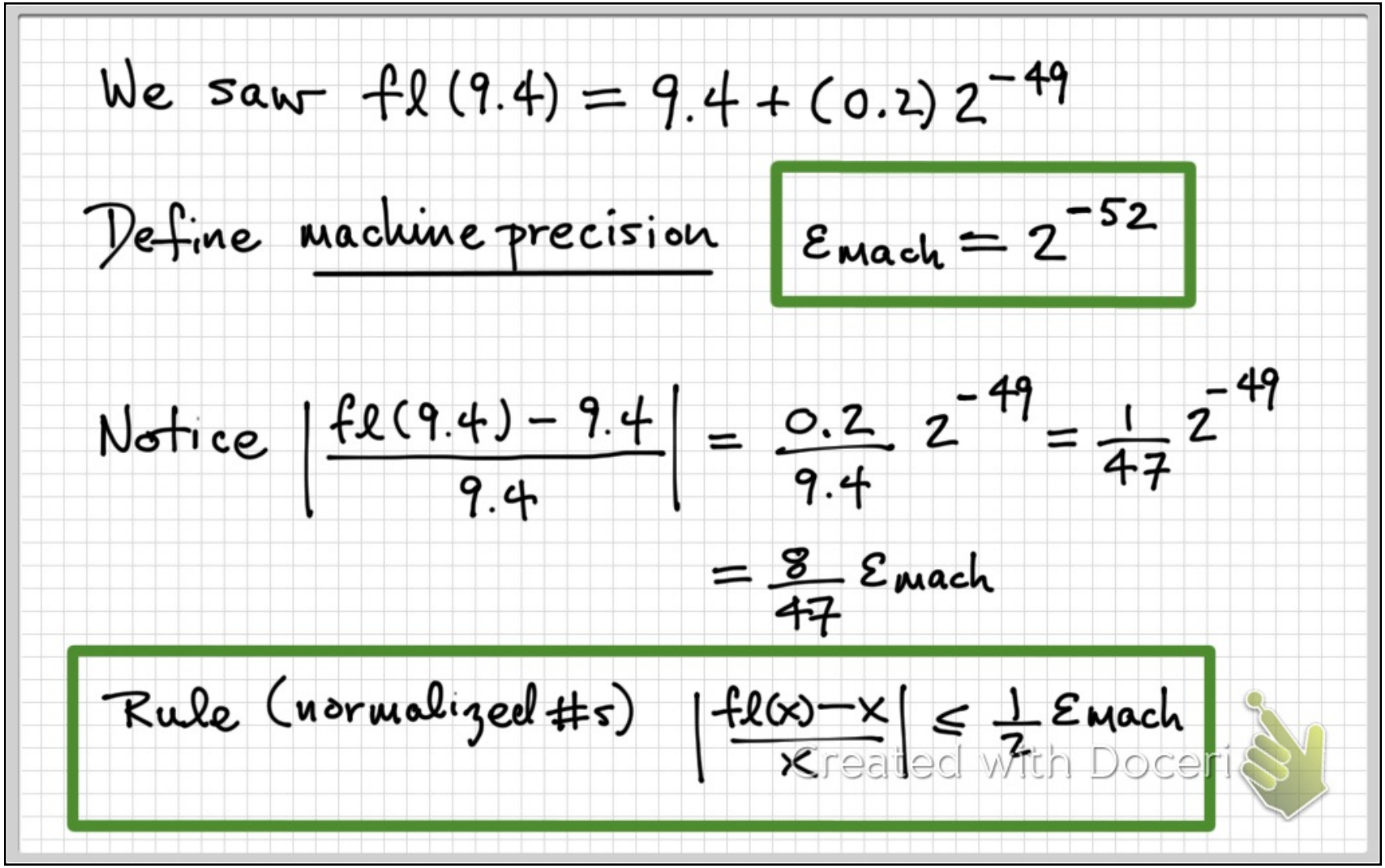
$= (1.001\ 0110\ 0110\ 0110\ 0110$
$0110\ 0110\ 0110\ 0110$
$0110\ 0110\ 0110\ 0110\ 0\,|\,110\ldots) \times 2^3$

↑ 52ND bit

$fl(9.4) = (1.001\ 0110\ 0110\ 0110\ 0110$
$0110\ 0110\ 0110\ 0110$
$0110\ 0110\ 0110\ 0110\ 1) \times 2^3$

We ① chopped $(0.110\overline{0110})2^{-52} \cdot 2^3$ & ② added $2^{-52} \cdot 2^3$


So $fl(9.4) = 9.4 - (0.1100\overline{0110}) \times 2^{-52} \cdot 2^3 + 2^{-52} \cdot 2^3$
a
b
a $= (0.\overline{0110}) \times 2^{-52} \cdot 2^4 = (0.4)\ 2^{-48}$
b $= 2^{-49}$
$fl(9.4) = 9.4 - (0.4)2^{-48} + 2^{-49}$
$= 9.4 + (0.2)2^{-49}$
Created with Doceri


We saw $fl(9.4) = 9.4 + (0.2)2^{-49}$
Define machine precision
$\varepsilon_{mach} = 2^{-52}$
Notice $\left|\frac{fl(9.4) - 9.4}{9.4}\right| = \frac{0.2}{9.4} 2^{-49} = \frac{1}{47} 2^{-49}$
$= \frac{8}{47} \varepsilon_{mach}$
Rule (normalized #5) $\left|\frac{fl(x) - x}{x}\right| \leq \frac{1}{2} \varepsilon_{mach}$
Created with Doceri

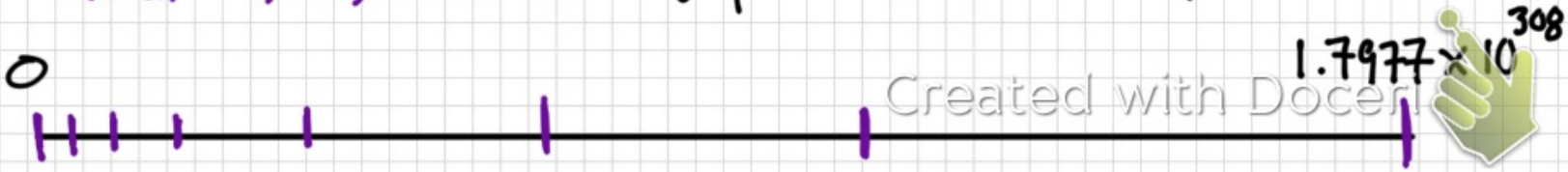
$\varepsilon_{mach} = 2^{-52} \simeq 2.22 \times 10^{-16}$, so $\frac{1}{2}\varepsilon_{mach} \simeq 1.11 \times 10^{-16}$

$$\left| \frac{fl(x) - x}{x} \right| \leq \frac{1}{2}\varepsilon_{mach} \simeq 1.11 \times 10^{-16}$$

near 1 $= 10^0$ gaps between floats about 10^{-16}

near 1000 $= 10^3$ gaps between floats about 10^{-13}

near 1,000,000 $= 10^6$ gaps between floats about 10^{-10}



Note: ε_{mach} is the smallest positive # such that

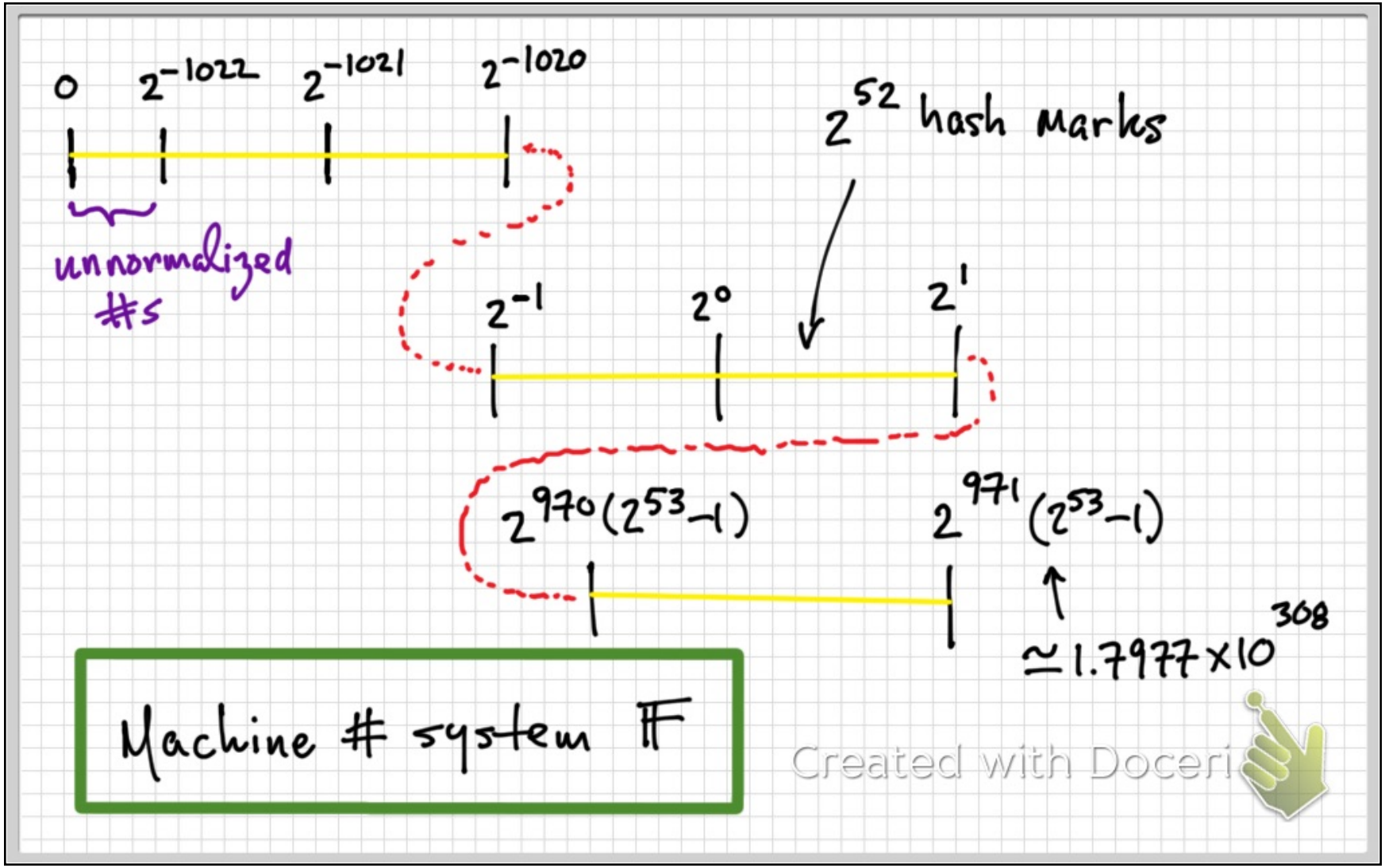
$$fl(1+\varepsilon_{mach}) \neq 1.$$

$1 + \varepsilon_{mach}$

$= (1.000\ldots0)\times 2^0$ ← 52ND bit

$+ (0.000\ldots1)\times 2^0$

$= (1.000\ldots1)\times 2^0$


0
2^{-1022}
2^{-1021}
2^{-1020}
unnormalized #s
2^{52} hash marks
2^{-1}
2^{0}
2^{1}
$2^{970}(2^{53}-1)$
$2^{971}(2^{53}-1)$
$\simeq 1.7977 \times 10^{308}$
Machine # system F
Created with Doceri

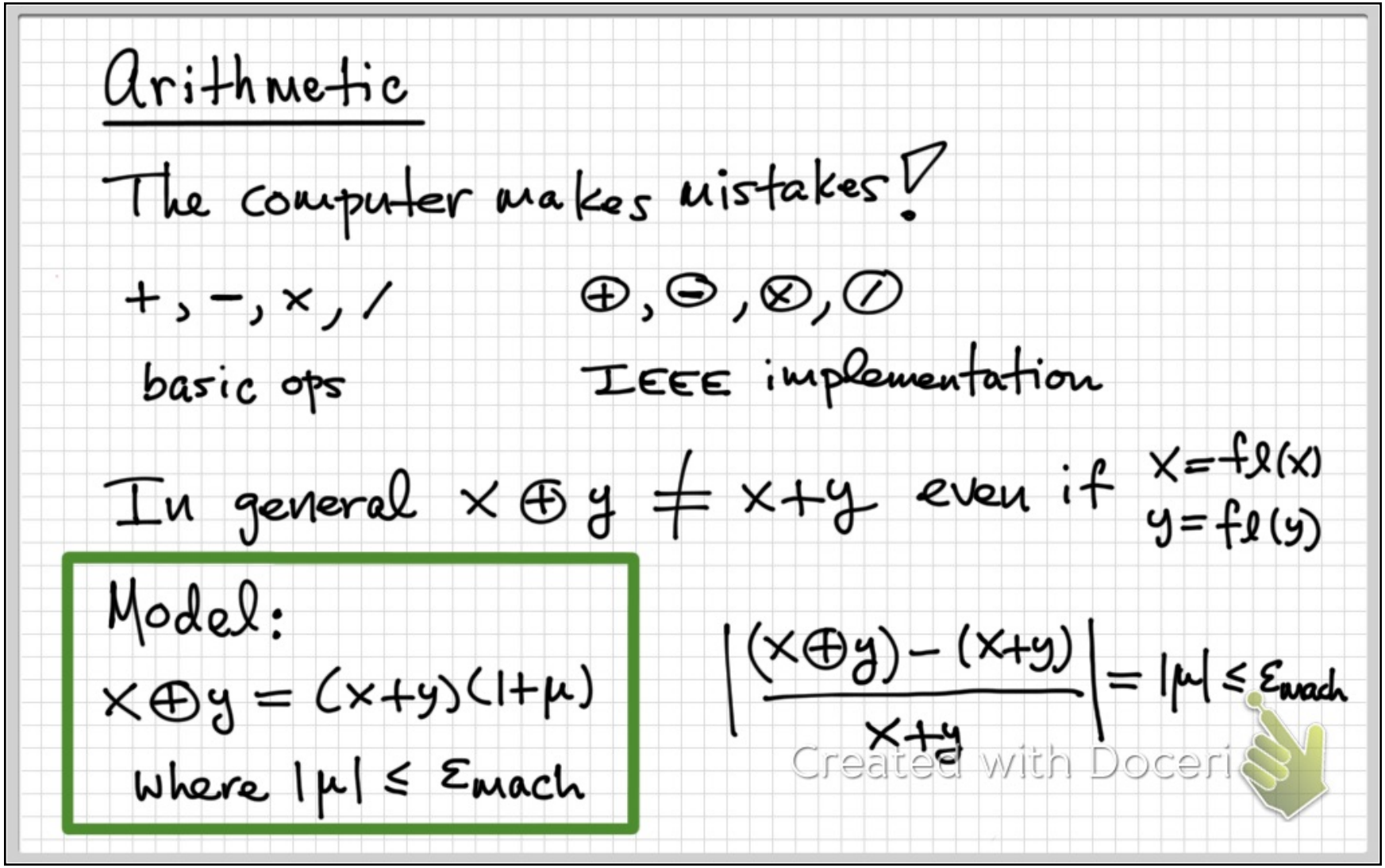
$$V_{max,norm} = (1.\underbrace{111\cdots1}_{52\text{ bits}})_2 \times 2^{2046-1023}$$

Notice $(1.11\cdots1)_2 + \underbrace{(0.00\cdots1)_2}_{2^{-52}}$

$= (10.00\cdots0)_2$

$= 2 \Rightarrow$ $\boxed{(1.11\cdots1)_2 = 2 - 2^{-52} = (2^{53}-1)\,2^{-52}}$

$V_{max,norm} = (2^{53}-1)2^{-52}\,2^{1023} = (2^{53}-1)2^{971}$


Arithmetic
The computer makes mistakes!
+, −, ×, /
basic ops
⊕, ⊖, ⊗, ⊘
IEEE implementation
In general x ⊕ y ≠ x + y even if x = fl(x), y = fl(y)
Model:
x ⊕ y = (x + y)(1 + μ)
where |μ| ≤ εmach
|((x ⊕ y) − (x + y)) / (x + y)| = |μ| ≤ εmach
Created with Doceri