CS2318

11/10/16

Real Number Representing

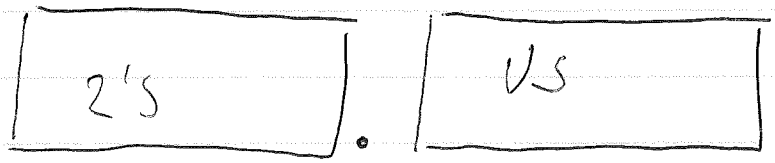
Imagining fixed point

Inside / to the left of a token

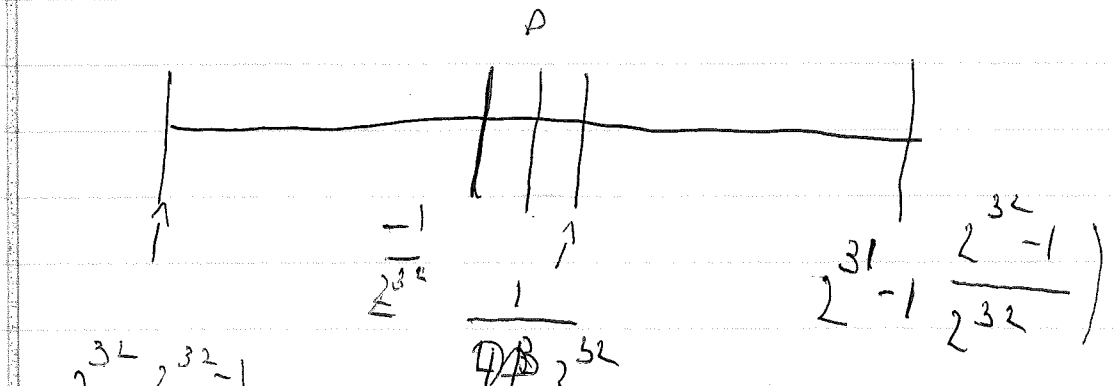
Division in MIPS

HI 2's

LO Unsigned



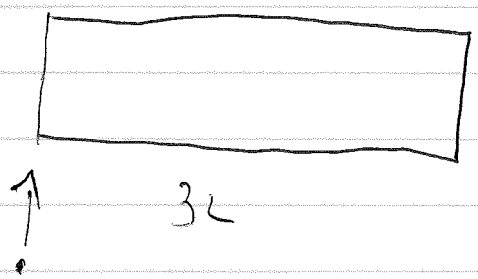
Int Remainder Inv



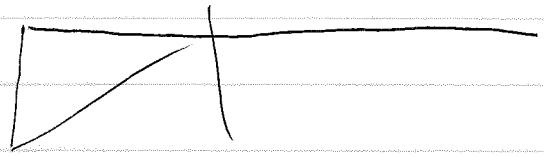
$-2^{32} \quad \frac{2^{32}-1}{2^{32}}$

$-(2^{32}+1$

2



32 $[-2^{31}, 2^{31}-1]$



$\left[-\frac{2^{31}}{2^{31}}, \frac{2^{31}-1}{2^{31}}\right]$ $(-1, 1)$

~~As~~ -1×-1 does not work

with the Int unit

$\times, \div, +, -$ of if these Fixed

Points is done with the Integer

Unit

(3)

IEEE 754 Floating Point

Standard

n bits code => real number

Real number => code

3.55×10^4 F.P.

Pseudo IEEE 16 bit (10) Used by HP

Float 32 - bit

Double 64 - bit

Quad 128 - bit

$35.5 \times 10^3 = \boxed{3.55 \times 10^4}$ Normal

$\boxed{0.355 \times 10^5}$ standard

only 1 Representation of a real #

(4) A 3.55×10^4

B $+ 305 \times 10^3$ Align exponent

move the decimal point to the left is equivalent to shifting the number point to the right $\Leftrightarrow$ shift left

$305.0 \times 10^3 = 30.5 \times 10^4$ B'

A 3.55×10^4 35.5×10^3

if $A > B$ $E_A > E_B$

Align to A $\Rightarrow$ Right shift (point to left) Lose LSB

5

A $\pm M_A \times 10^{E_A}$ 3.55×10^4

\+

B $\pm M_B \times 10^{E_B}$ 305.0×10^3

1) Normalize

$d.xxx \times 10^E$ $d \neq 0$ x any digit

$\Rightarrow 0.dxxx\ 10^E$ $d \neq 0$ x " "

A' 0.355×10^5

B' 0.305×10^6

2 Align Exponent

$E_{B'} > E_{A'}$ align to B

A'' $+0.0355 \times 10^6$ $S_A M_A E_A = (+) \pm M_A \times 10^{E_A}$

\+

B $+0.305 \times 10^6$ $S_B M_B E_B = \pm M_B \times 10^{E_B}$

C $E_{OA} =$ $S_C M_C E_C$

3 $E_C = E_{B'}$

6 $M_C = M_{A''} + M_{B'}$

$\pm (M_{A''} + M_{B'})_x$ $10^{E_{B'}}$ result

7 Normalize / Round ↺ potentially twice

8 check for overflo

F. P. Multiplication

3.55×10^4 $+0.355 \times 10^5$

$\times$ $*$

305.0×10^3 $+0.305 \times 10^6$

$S_A M_A E_A$

$*$ $= S_C M_C E_C$

$S_B M_B E_B$

$S_C = S_A \oplus S_B$

1) Normalize

2) $S_C = S_A \oplus S_B$

3) $E_C = E_A + E_B$ (check for ovf)

7

4) $M_C = M_A * M_B$ (2 ints)

5) Normalize Round

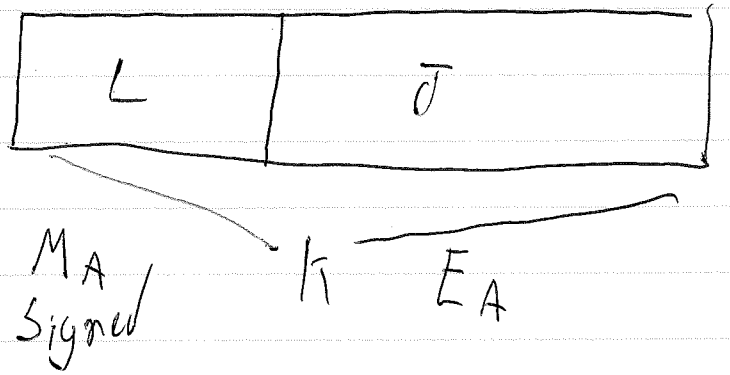
6) check for OVF

1) Naive Representation $B \times 10^{-3}$

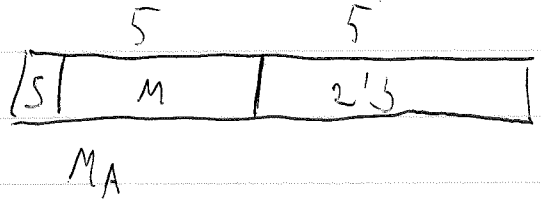
=> improve Naive

a. Efficient coding

b. Reuse (of existing units)

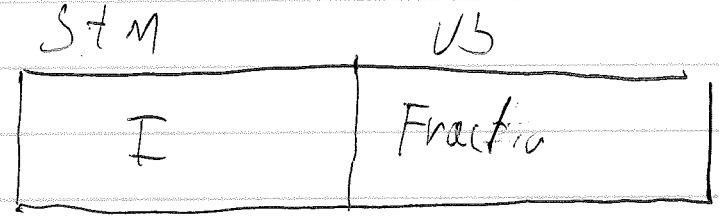


Assume S+M 2's



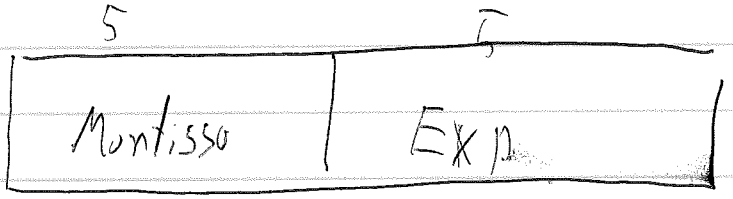
8

A



5 ↑ 5

B



5+5 2'5

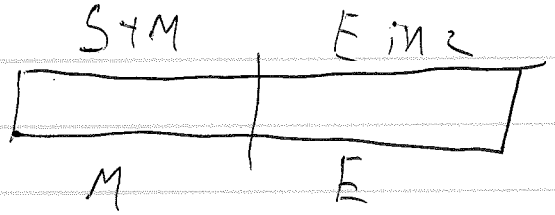
S+M 2'5

11110 | 10111

A $-14\frac{23}{32}$

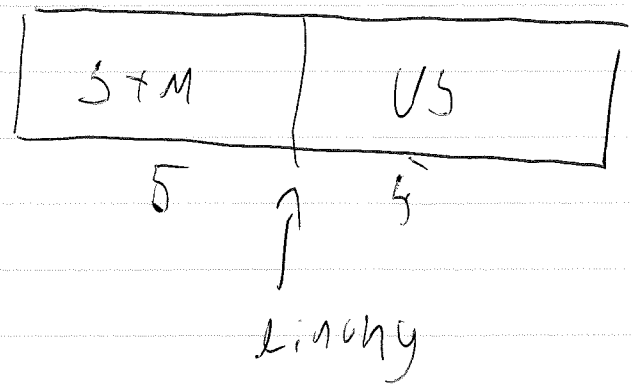
B -14×2^{-9}

$(S)^{-1} \times M \times 2^{E}$ Novik

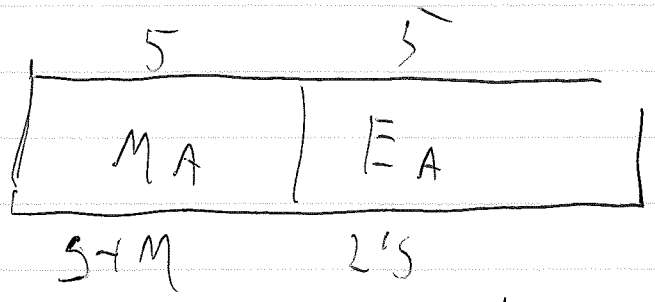


10111 01001

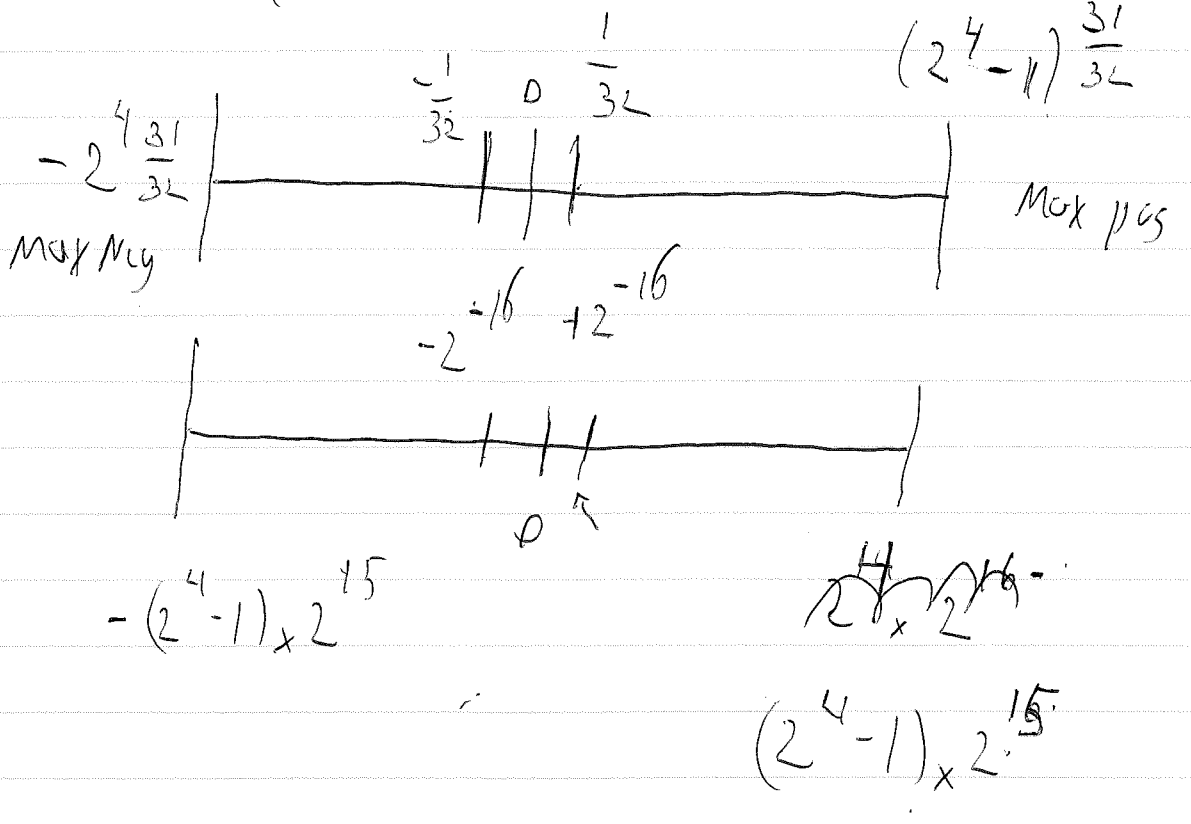
9



Fixed



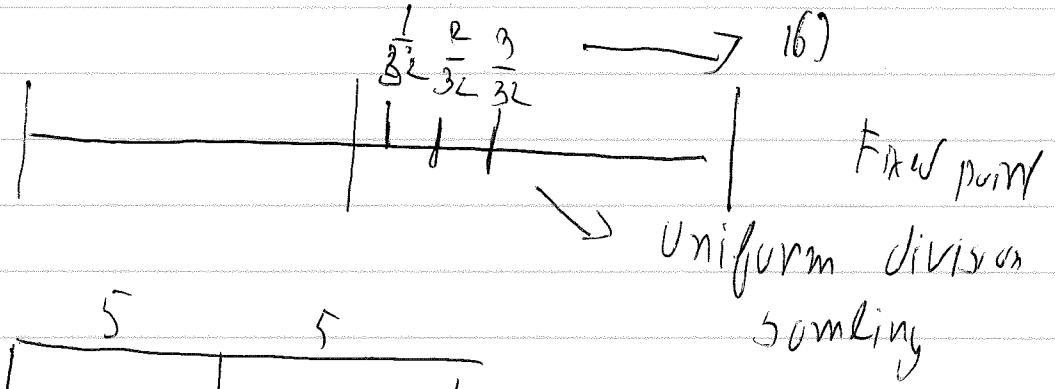
Float

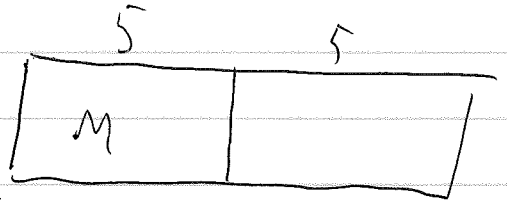


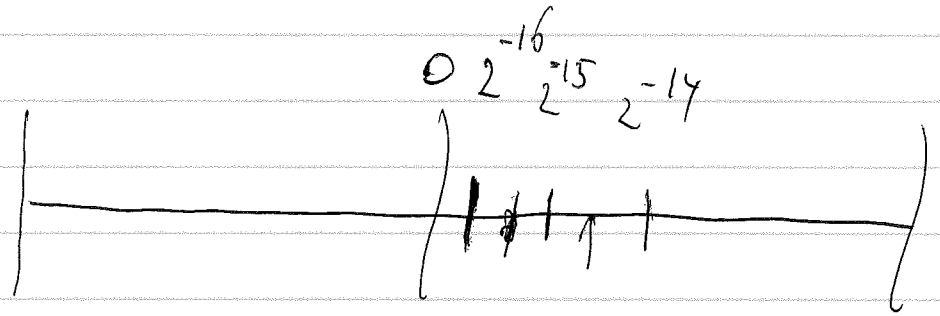
Why Fixed point

Why Float Large dynamic range

10





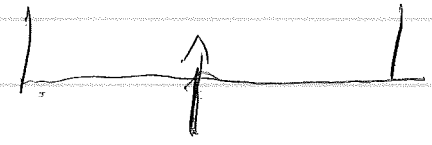


Exponential Logorithmic division

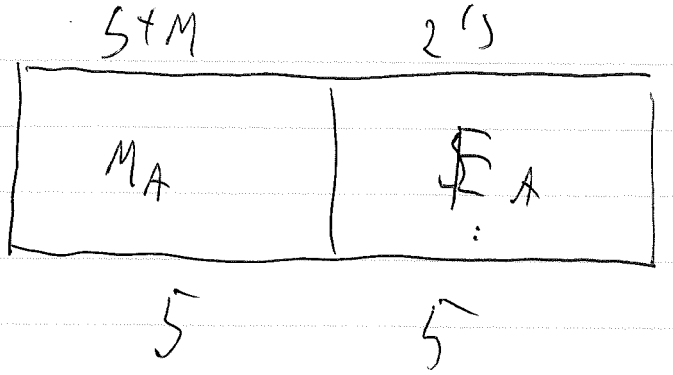
Fixed

With Fixed worst error is $\frac{1}{2}$ LSB

5 bits $\frac{1}{2}$ LSB $\frac{1}{64}$ $\frac{1}{2^5}/2$



14



2^6 Representation

$\sim$ 15 Representation

NORMALISE