

3-1/

$0 \ 1000 \ 0001 \ 1100 \ 0 \text{---} \quad (7)$   
 $0 \ 1000 \ 0010 \ 0 \text{---} \quad (8)$   
 $0 \ 1000 \ 0010 \ 1110 \text{---} \quad (15)$   
 $\underbrace{130-127=3} \quad (1+1/2+1/4+1/8) \times 2^3$

$\langle +, 2, 1.75 \rangle \quad (7)$   
 $\langle +, 2, 1 \rangle \quad (4)$   
 $\langle +, 2, 2.75 \rangle = 11$   
 $\hookrightarrow \langle +, 3, \frac{2.75}{2} = 1.375 \rangle = 11$

$1201600 \ 0101 \ 11100 \text{---}$   
 $+ 0 \ 1000 \ 0010 \ 10110 \text{---}$   


---

 $6 \quad 11110 \ 000$   
 $3 \quad 0 \ 0011 \ 011$   


---

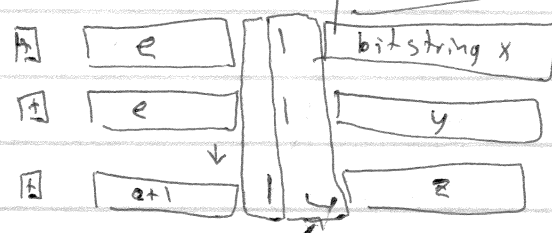
 $7 \quad 10 \ 0001 \ 011$   


---

 $1000 \ 0110 \ 0001 \ 0110 \text{---}$

$\langle +, 3, 1 \rangle$   
 $\langle +, X, Y \rangle$   
 $+ \langle +, X, Z \rangle$   
 $= \langle +, X, \square \rangle$

$0 \ 1000 \ 0110 \ 0000 \ 1011 \ 0 \text{---}$   
 $2^7 \quad 1.04296875$



$x+y = z$   
 $\underbrace{\quad}_{23\text{-bit}} \quad \underbrace{\quad}_{23\text{-bit}} \quad \underbrace{\quad}_{24\text{-bit}}$

$\langle +, 2, 1.75 \rangle$   
 $7 \quad 0 \ 1000 \ 0001 \ 1100 \text{---}$   
 $8 \quad 0 \ 1000 \ 0010 \ 0 \text{---}$   
 make 8 smaller  
 $\langle +, 2, 2.00 \rangle$

make 7 bigger

$7 \quad 0 \ 1000 \ 0010 \ 0 \ 1100 \text{---}$   
 $\langle +, 3, 0.875 \rangle$   
 $2^{-23} \times 2^e$

$\frac{2^7}{2^8} \times 1.75 = 2^3 \times X$

Adding loses information about smaller number

$\langle +, e_x, m_x \rangle$   
 $+ \langle +, e_y, m_y \rangle$  (assume  $e_y > e_x$ )  


---

 $\langle +, e_y, m_z \rangle \quad m_z = m_y + (m_x \gg (e_y - e_x))$   
 $e$  field is 8-bits  $\in [-127, +127]$

$$\begin{array}{r} \frac{e=100 \quad (x)}{+ e=25 \quad (y)} \\ \hline = x \end{array}$$

loss: 75-bits  
my 24-bits

for (float f=10mil; f>0, f-=1.0) {  
} ... }

Algorithm for Addition: — Sort numbers ~~for~~ X and y  $\Rightarrow$  small and big

— Turn mantissa(s) into 24-bit numbers (orig 23-bit)

— Right-shift  $M_{small}$  by  $(e_{big} - e_{small})$

— Add mantissas  $\Rightarrow$  25 bits (24 + 24 = 25)

25 24 mantissa

00  $\rightarrow$  not possible

01  $\rightarrow$  easy (common case of large + small)

10  $\rightarrow$  increment e, right-shift by 1 (carry)

11

$\rightarrow$  lost  $2^{-23} \times 2^e$

Floats  $\neq$  Reals

$\langle +, -, \cdot, / \rangle$

7.000 —

$2^{32}$

$$x = a + b + c;$$

$\Rightarrow$

$$t = b + c;$$

$$y = b + c + d;$$

$$x = a + t;$$

$$y = t + d;$$

b and c are constant  
for (changes a)  
for (changes d)

Reals are associative

$\Rightarrow$

$$(x + y) + z = x + (y + z)$$

$$x(y + z) = xy + xz$$

Floats are NOT

Floats don't

"fast-math"

- ffast-math

$$a \geq b$$

$$a + x \geq b + x$$

(monotonicity) [ ~~$\neq$~~   ~~$a \rightarrow a$~~   $a + 0 = a$ ]

Round-to-even  $\Rightarrow$  make the number even (stats wanted)

Round-to-zero  $\Rightarrow$  closer to zero

Round-down  $\Rightarrow$  ~~truncate~~ truncate

Round-up  $\Rightarrow$  ceiling