Note Title 11-09-2012

Sep-11

1) Floating point addition, multiplication.

$1.5 + 0.01$

$$\begin{array}{r} 1.50 \\ +\ 0.01 \\ \hline 1.51 \end{array} \qquad \begin{array}{r} 1.59 \\ +0.01 \\ \hline 1.60 \end{array}$$

decimal

1) align the decimal points

2) Perform the addition.

$$\begin{array}{r} 5.5 \\ +\,5.5 \\ \hline 11.00 \end{array}$$

Restrictions:

i) The number of digits to the left of the decimal point is 1

ii) Number of digits to the right of the point is limited to p.

$p=1$

$$\begin{array}{r} 5.5 \\ +5.5 \\ \hline 11.00 \end{array} = 1.1 \times 10^1$$

3) If there is a carry out, right shift and add 1 to exponent.

$p=1$

$5.5 + 5.5 \times 10^{-1}$

$$\begin{array}{r} 5.5 \\ + .55 \\ \hline 6.05 \end{array}$$

$\underbrace{05}_{2 \text{ bits}}$

4) Round

$\rightarrow 6.0$ Truncate

↳ 6.1 Increment

$5.4 + 5.9 \times 10^{-1}$

$$\begin{array}{r} 5.4 \\ +\ .59 \\ \hline 5.99 \end{array}$$

$9.4 + 5.9 \times 10^{-1}$

$$\begin{array}{r} 9.4 \\ +\ .59 \\ \hline 9.99 \end{array}$$ (round)

10.0

5) If rounding leads to a carry out goto step 3

$p=1$

IEEE 754 - Methods of rounding.

	3.55	-3.55
$+\infty$	3.6	-3.5
$-\infty$	3.5	-3.6
0	3.5	-3.5
nearest (even)	3.6	-3.6

Binary FP addition.

$(n_1 \geq 0,\ n_2 \geq 0)$

Two numbers : n_1 n_2 (assume
(e_1) (e_2) $n_1 \geq n_2$)

1) Align decimal points
   i) Shift n_2 $(e_1 - e_2)$ positions to the right

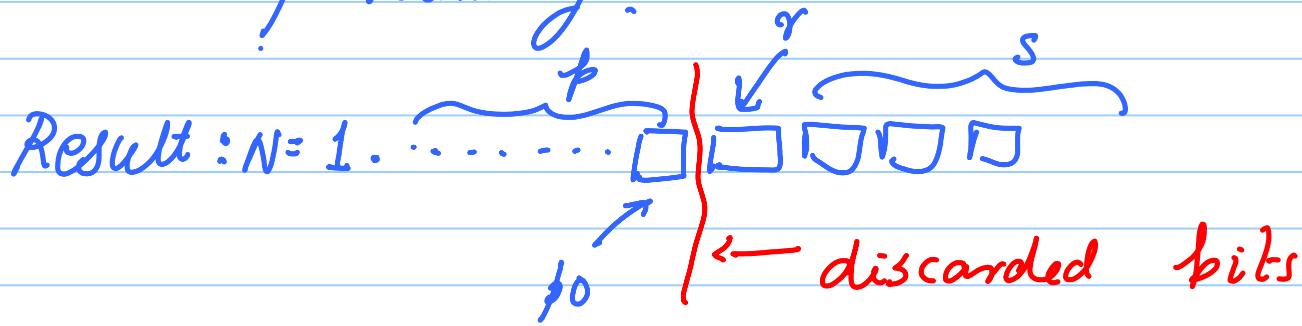
2) Perform addition

3) Adjust for carry-out

4) Rounding

5) If rounding leads to carry out - goto step 3

How to do binary rounding?



$p_0 \rightarrow$ LSB of the mantissa

r (round bit) $\rightarrow$ MSB of discarded bits.

S (sticky bit) $\rightarrow$ OR of the rest of the discarded bits

Example.

$p=2$

$1.\overbrace{01}0110$

$p_0 = 1$

$r = 0$

$s = 1$

$1.\overbrace{01}^{p}\overset{r}{1}\overbrace{0000}^{V}$

$p_0 = 1$

$r = 1$

$s = 0$

$N = 1.\overbrace{\cdots p \cdots}\ \|\ \overbrace{\cdots\cdots}^{D}$

$N = 1.\overbrace{------}^{p} + \Delta$

$= N_0 + \Delta$

$(r \wedge s) \Rightarrow (\Delta > 0.5 \times 2^{-p})$

$\text{Truncate}(N) = N_0$

$\text{Inc}(N) = N_0 + 2^{-p}$

$(x>0)\ [(r \vee s) \not\Rightarrow (\Delta > 0)]$

$1.\overbrace{-\,-\,-\,-\,-\,-}^{p}\,\big|\,\boxed{r}\,\boxed{s}$

p_0

Action Table

[If you are not incrementing, you are truncating]

	$x \geq 0$	$x < 0$
$+\infty$	$(r \vee s) \Rightarrow Inc$	
$-\infty$		$(r \vee s) \Rightarrow Inc$
0		
nearest (even)	$[(r \wedge s) \vee (r \wedge p_0)] \Rightarrow inc$	$[(r \wedge s) \vee (r \wedge p_0)] \Rightarrow inc$

Same idea for multiplication.

Step 1 and 2 differ

1. Set the sign bit, (add exponents, − bias)

2. Perform multiplication.

3
4
5 Same