Note Title Sept -16 16-09-2011

Floating point operations

Addition

$10.764 + 2.3$

$$\begin{array}{r} 10.764 \\ +\ 2.3 \\ \hline 13.064 \end{array}$$

1) Align the decimal points.

$2.3\times10^2 \quad + \quad 4.6$

$= (0.046 \times 10^2)$

$$\begin{array}{r} 2.3 \\ +\,0.046 \\ \hline 2.346 \end{array}$$

1) Align Decimal Points

a) Shift the smaller number to the right s.t both the exponents match

$\underset{A}{(1.x)} + \underset{B}{\begin{pmatrix} 1.y \\ 0.z \end{pmatrix}}^{\text{or}}$

2) ·

$A+B$

$\left.\text{(Perform the addition}\right\}$ [Maintain some extra precision beyond 23 bits]

3) $(1.1)_b + (1.1)_b = (11.0)_b$

$\nearrow$ Not Normalized

Renormalize the number.

$(11)_b = (1.1 \times 2^{1})_b$

4) At the moment, number is

$$\left[1.(\underbrace{x_1 - - - - - x_{23}}_{23 \text{ bits}}) \times 2^{E} \right]$$

$$1.\ \underbrace{x_1 \ . - - - \ x_{23}} \ \underset{(x_{24} - - - -)}{(\underline{\qquad\qquad})}$$

Discard digits (x_{24}) onwards

Try to round x_{23}

Example:

$$1.\underbrace{(1\ 1\ 1\ \cdots\ 1)}_{23}\ \underbrace{\overset{\times}{(1\ 1\ 1\ 1)}}_{4}$$

$$= (10.0)_b$$

$$= (1.0)_b \times 2^1$$

5) After rounding – my number might get un-normalized.

Renormalize.

Rounding

Assume that you want to round to the next integer.

4.5 =

4.4 = 4

4.6 = 5

Rounding Policies:

Round-Up (+∞) : 4.1 → 5

Round-Down (−∞) : 4.9 → 4

Truncate : 5.369 → 5

Round (complicated)

(+)ve Trunc = Round-Down

(-)ve Trunc = Round-up

1.0 .... 1.5 → 1

1.50001 .... 1.9999 → 2

<u>**Assume**</u>

Implement $\overbrace{\text{round to next integer}}^{\text{complicated}}$ using the complicated rounding scheme.

$$N = \{1.r\underbrace{x_1 \, - - - - \, x_n}_{n}\}_b$$

$(r==0)$ $N = \underline{1}$

$(r==1)$ $\{$ $S = x_1 | x_2 \, - - - \, x_n$

$$1 . r(\overbrace{x_1, \dots x_n}^{(x_1 | x_2 \dots | x_n) = s})$$

round bit

sticky bit

```
if ($s==0$) {
    $N =. 1$
}
else if ($s==1$) {
    $N = (2)_d = (10)_b$
}
}
```

Result: $1.[x_1 - - - - - x_{23}][\underset{r}{\boxed{x_{24}}}\ \underbrace{x_{25} - - - x_{30}}_{\text{OR} \downarrow s}]$

We decide whether to increment

x_{23}

Extra

Addition

✓

Subtraction

1) Take a look at sign bits.

2) Do a sub or add

3) Set value of final sign bit

Multiplication

1) Figure out sign of the result

2) Add the exponents

Division

1) Sign of the result

2) Subtract exponents

Floating Point Tricks

$(y > 0)$

Is: $(x + y > x)$

(maybe)

$\{a + b - c \times d + e\}$

FP operations are not associative

$$\left.\begin{array}{r} 1.0000 \\ +1\times 2^{-50} \\ \hline 1.0000\ldots \end{array}\right\}$$ use double (52 bits)

$1 + 2^{-50}$