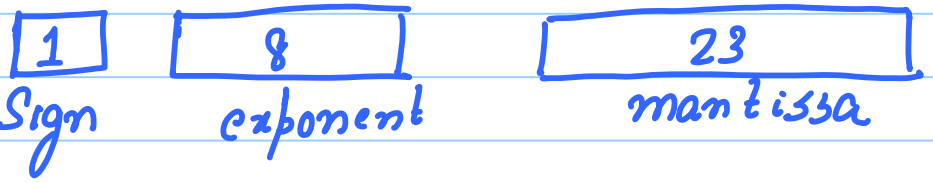
Note Title 13-09-2011

# Sept 13th

## Floating Point




0 (E=0, m=0)

+∞ (255(E), m=0)

−∞ (255(E), m=0)

Nan (255(E), m≠0)

Denormal (E=0) (m≠0)

$$\begin{array}{r} 7.834 \\ 3.212 \\ \hline 11.046 \end{array} \quad (1.1046 \times 10^1)$$

$(1.105 \times 10^1)$

Ex: 1

$+2^{-50}$

$x + y = x$

## Two Floating Point Numbers $(A+B)$ $(A>0 \ \& \ B>0)$

1) Take a look at sign bits
2) Compare A and B (W.l.g $A>B$)
3) Align the smaller number, B, by right shifting.
4) Add mantissas. $(1.x + 1.y = 2.u$ or $3.v)$

5) Normalize

6) Rounding

7) Re-normalize.

## Examples.

precision = 3 bits

$1.011$

$+$

$1.11 \times 2^{-2}$

$1.011$

$0.0111$ ← LSB

$1.1101$

1) Truncate

2) Round to $+\infty$

3) Round to $-\infty$

4) Round :

# Rounding an FP Num

FP → Int

1.---- 1.x $(0 < x \leq 0.5)$ → nothing

$\{0 \leq x < 0.5,\ x_0 = 0\}$ nothing.

$x_0\ x_1\ x_2 \underbrace{----}_{s}$ $(0.5 < x < 1)$ → increment

$x_0 \neq 0$

↳ $x = 0.5$ (nothing) $(s == 0)$

↳ $x > 0.5$ $(s == 1)$

Round to 1. 23 bits

r → 24th bit

s → 25...n ← OR.

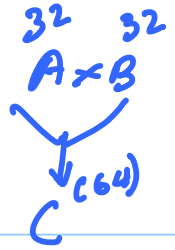
## Errata.

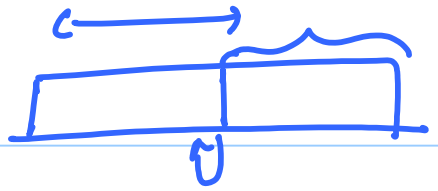
$x_0$ → round bit

$x_1 | x_2 | \ldots x_n$ = s → sticky bit.

$(r \& s)$ increment

## Multiplication




[Notes consider a guard & round bit (conceptually same)]

---

Truncate → Round down (+)
Round up (-).

Round up & Round down have the same connotation.

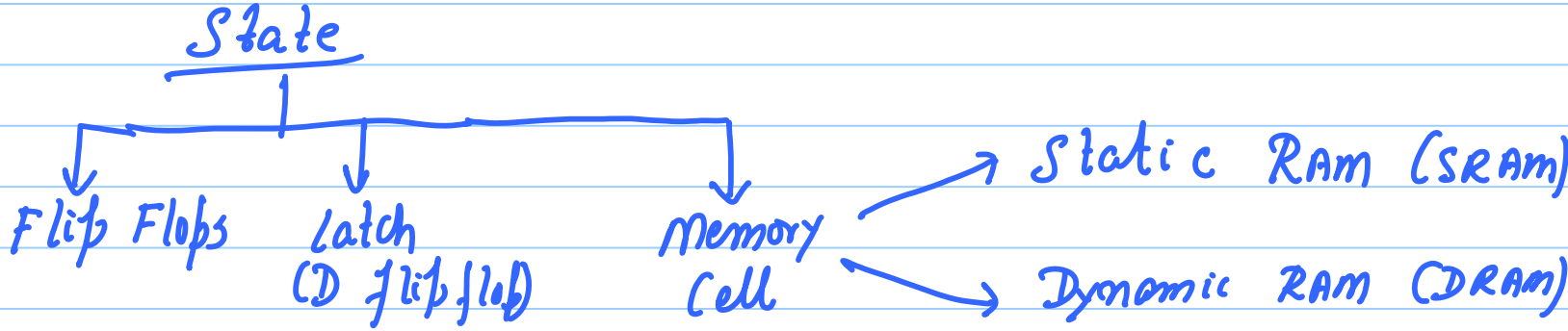
{ Floating Point Assembly Not covered }
Please Read.

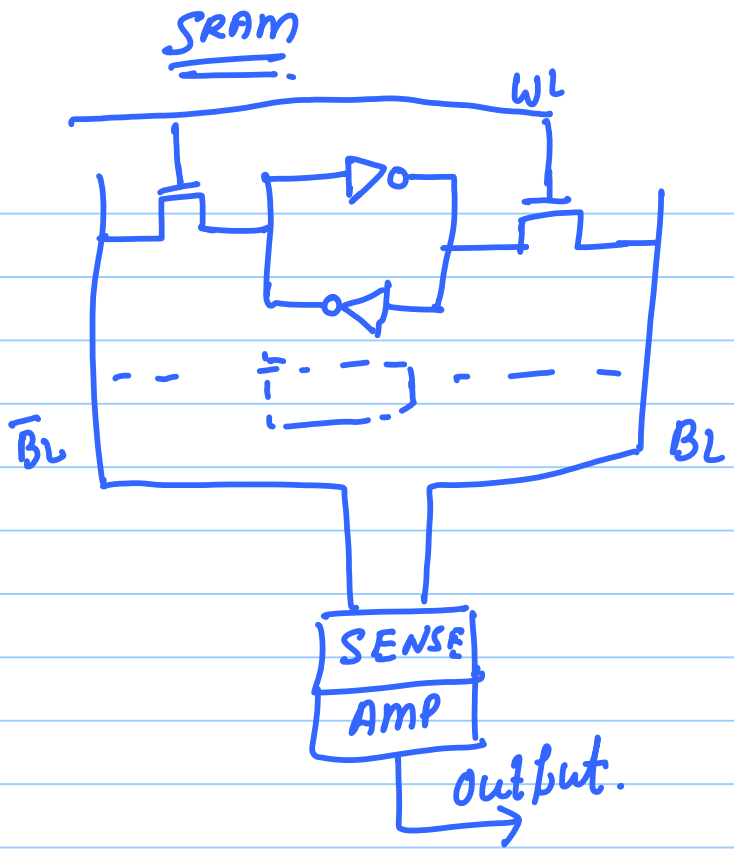
# Chapter 4

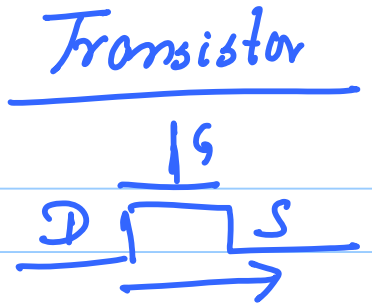
1
2
3 Architecture

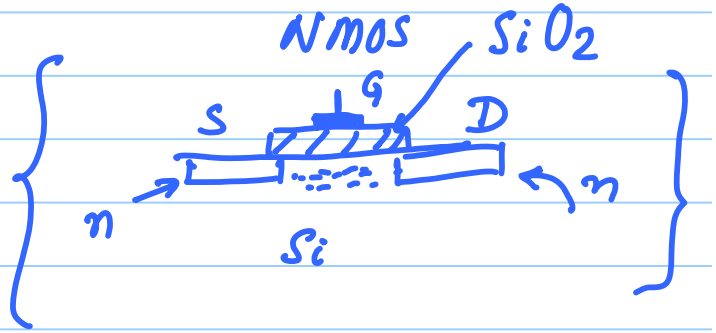
- → State (Registers, Mem.)
- → logic. (Adder, Sub, Divide, ...)

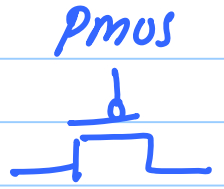
4
5
6 Organization.





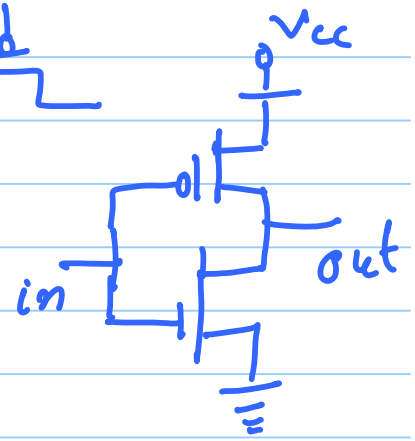




Once I enable the wordline.
BL or $\overline{BL}$ will swing to 1
other will swing to 0

Inverter



$V_{dd} = 1V$

1) Precharge: BL and $\overline{BL}$ to 0.5 V

2) Enable WL

3) BL and $\overline{BL}$ would start swinging in opp. directions.

4) 450 mV \ 550 mV