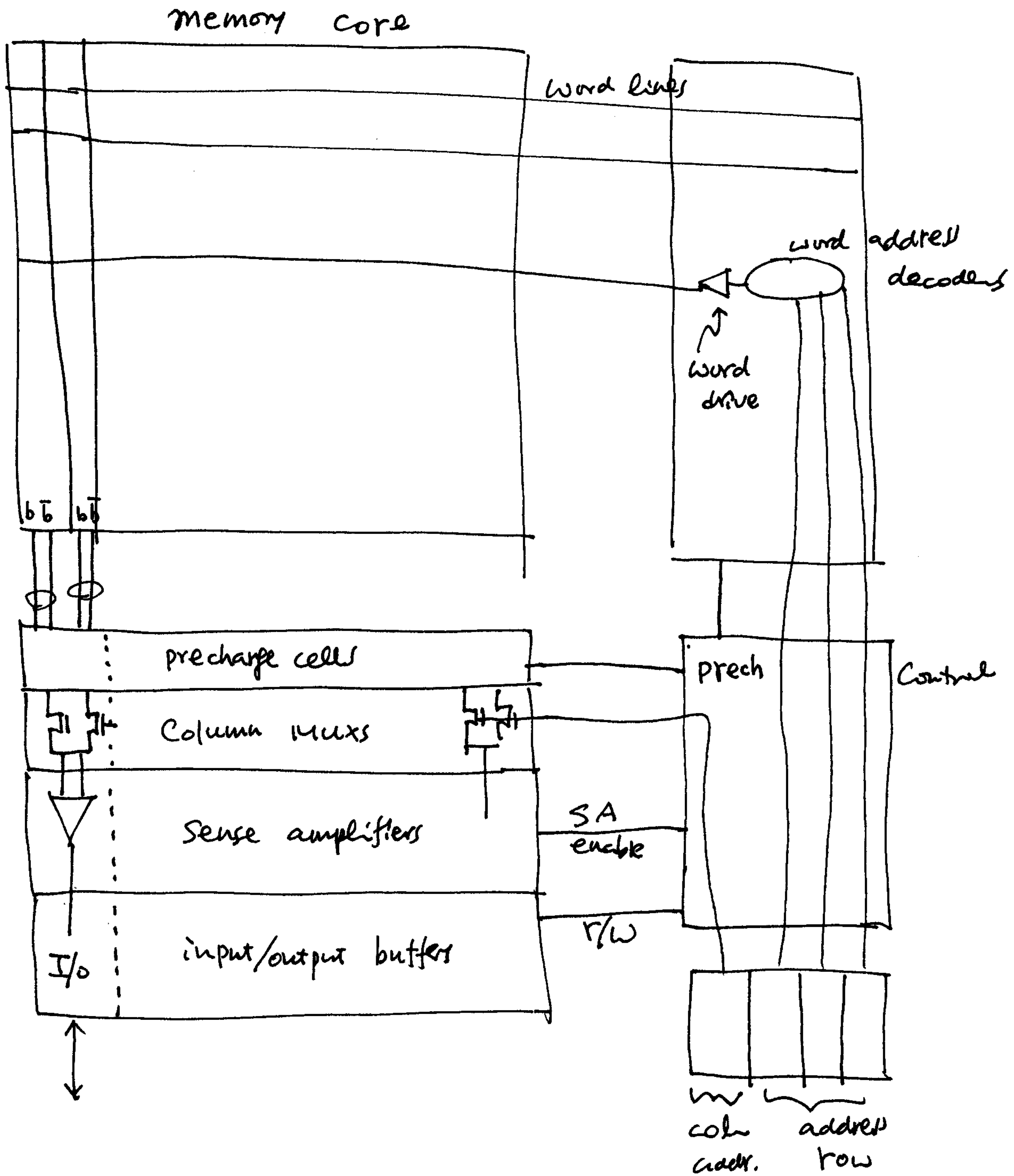
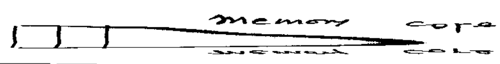
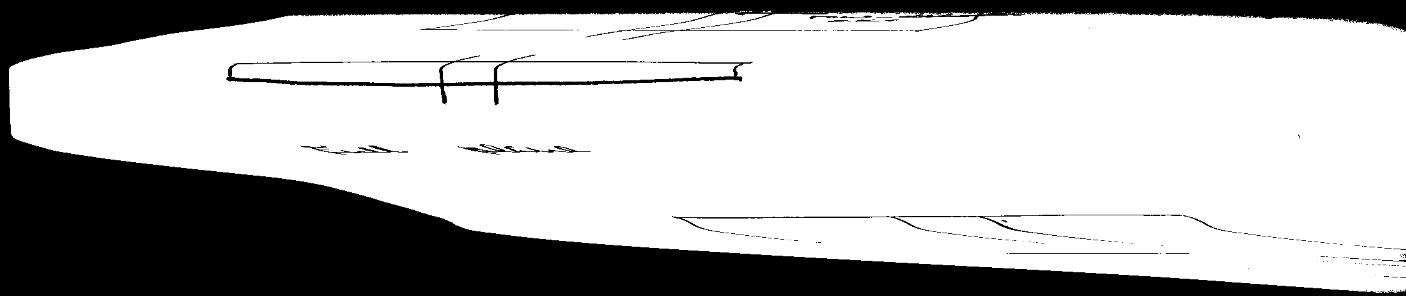
• Organization of a RAM.

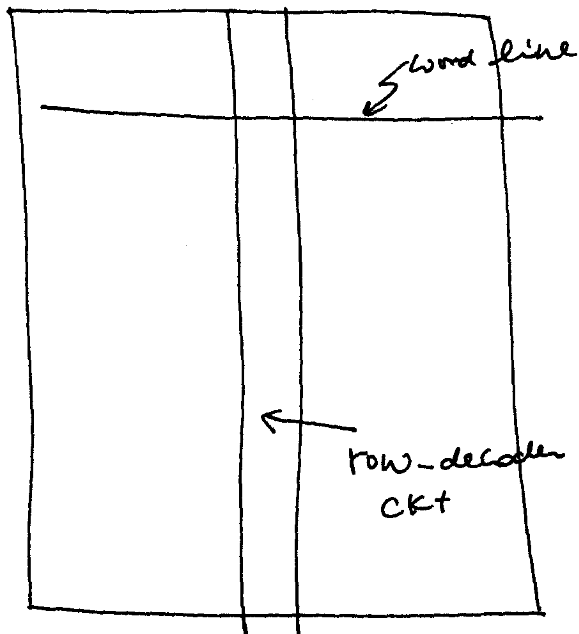


• Organization of a RAM.



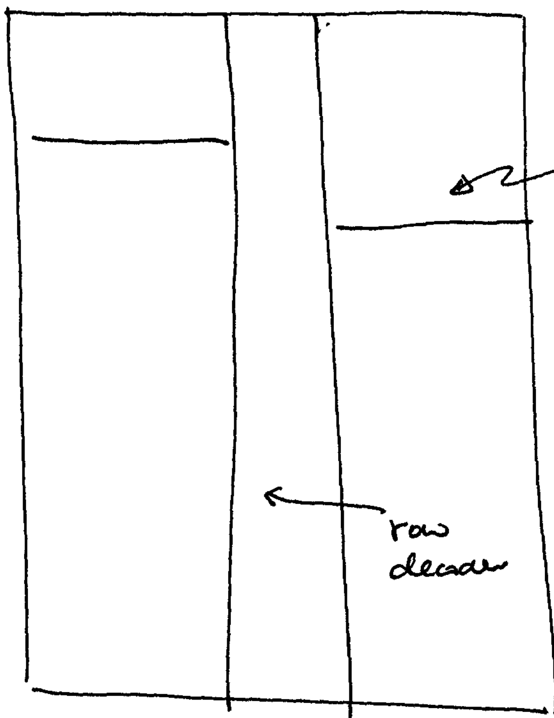


Memory Array Architecture



- Each row address accesses both sides of the array
- Good for not too large capacity array.

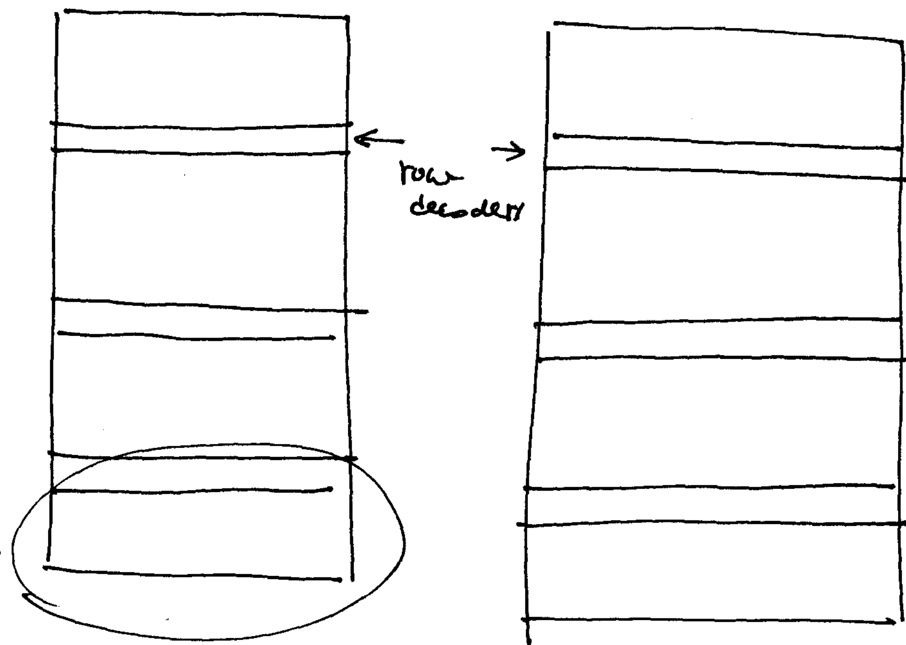
Full plane



- Each row address accesses one side of the array
- Lower power consumption.

Half plane

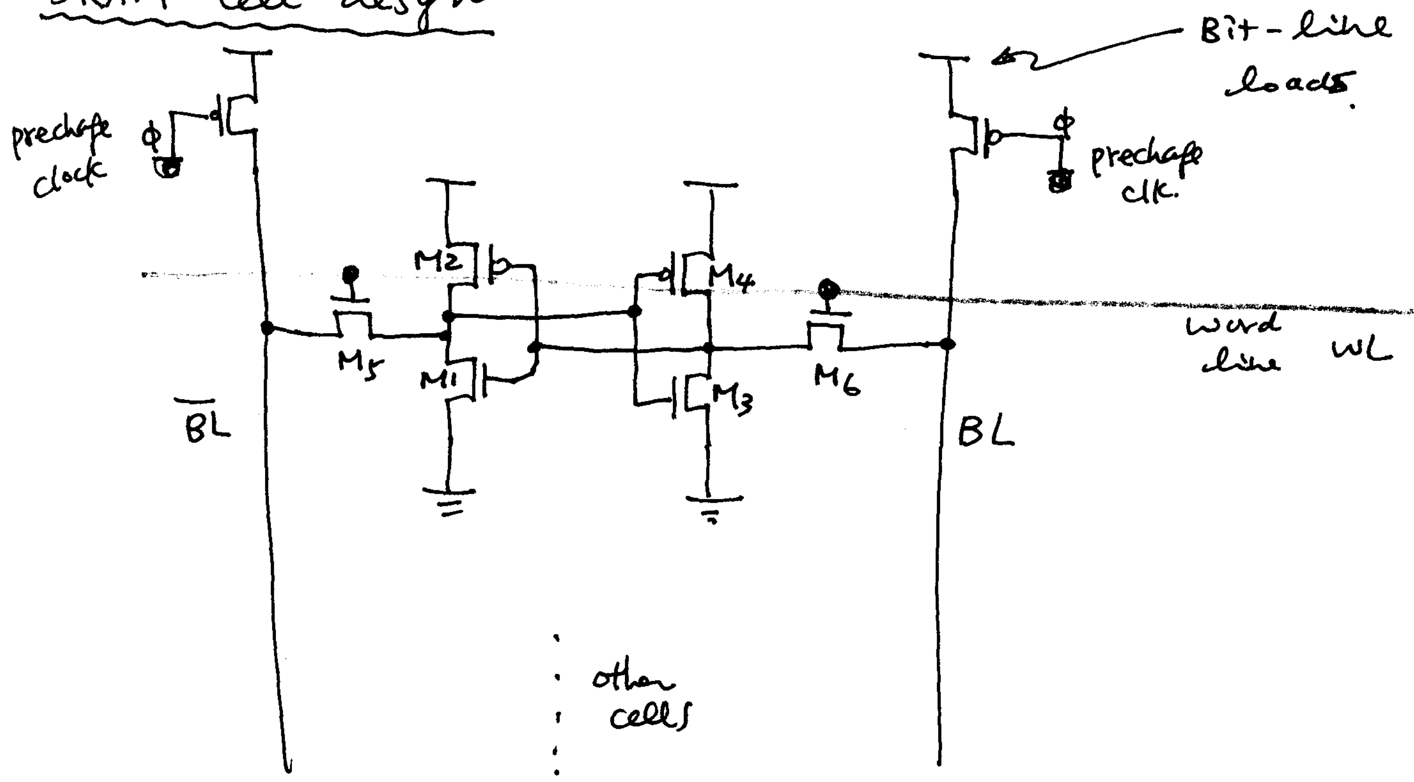
- Block-oriented



- Reduce bit line length

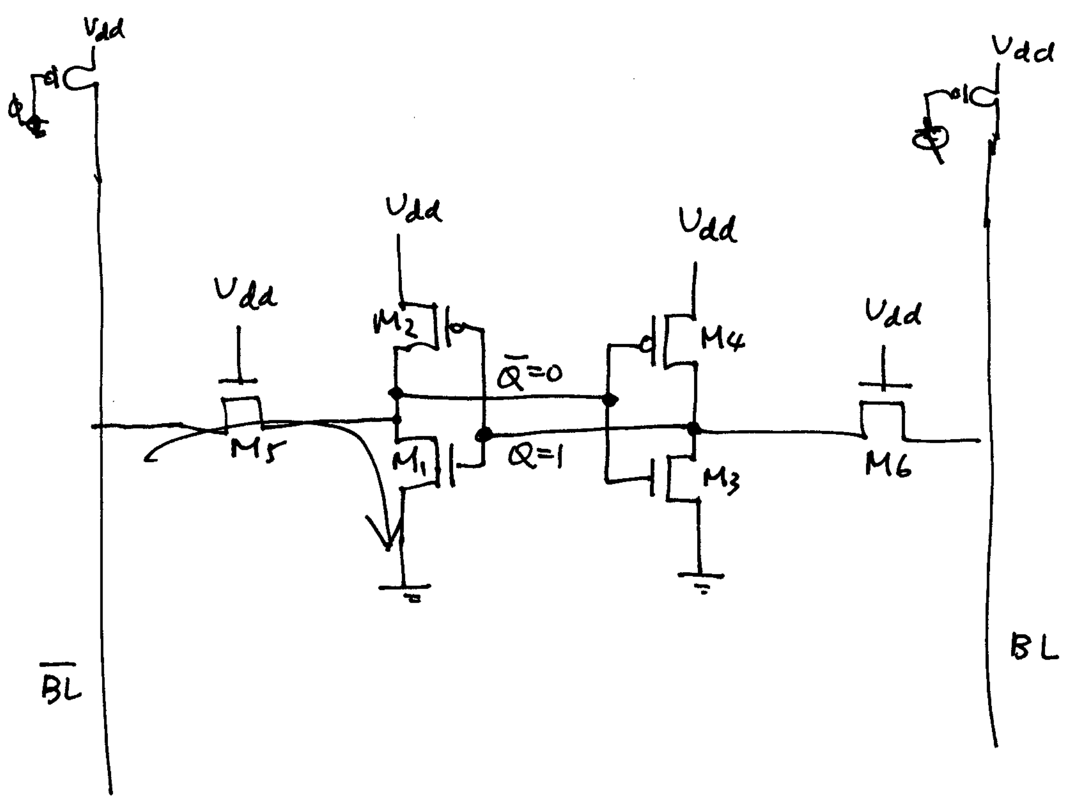
Each block contains its own Block selection (to select a specific block) ckt, x-decoder (select one of the rows) and sense amplifier.

- SRAM cell design



- Transistor sizing is very important.

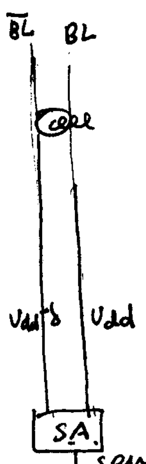
- Exaple for Read



- Assume the cell stores logic 1

  ∴ $Q=1$, $\bar{Q}=0$

- Before read, both bit lines are precharged to 1.



- We expect $BL=1$, but $\overline{BL}$ will be pulled-down

- Capacitances of $\overline{BL}$ and BL are much larger (pF) than transistor capacitances, so the swing of $\overline{BL}$ will not be too large

  ⇒ needs a sensor amplifier to amplify the signal.

- $\overline{Q}=0$ cannot be ~~dst~~ destroyed.

i.e., the voltage of $\overline{Q}$ can not be pulled up too much.

$$V_{\overline{Q}} = \left(\frac{V_{dd}}{R_{M5} + R_{M1}}\right) R_{M1} \approx 0.3\,V_{dd}$$

3.5) 1.00 → 0.3

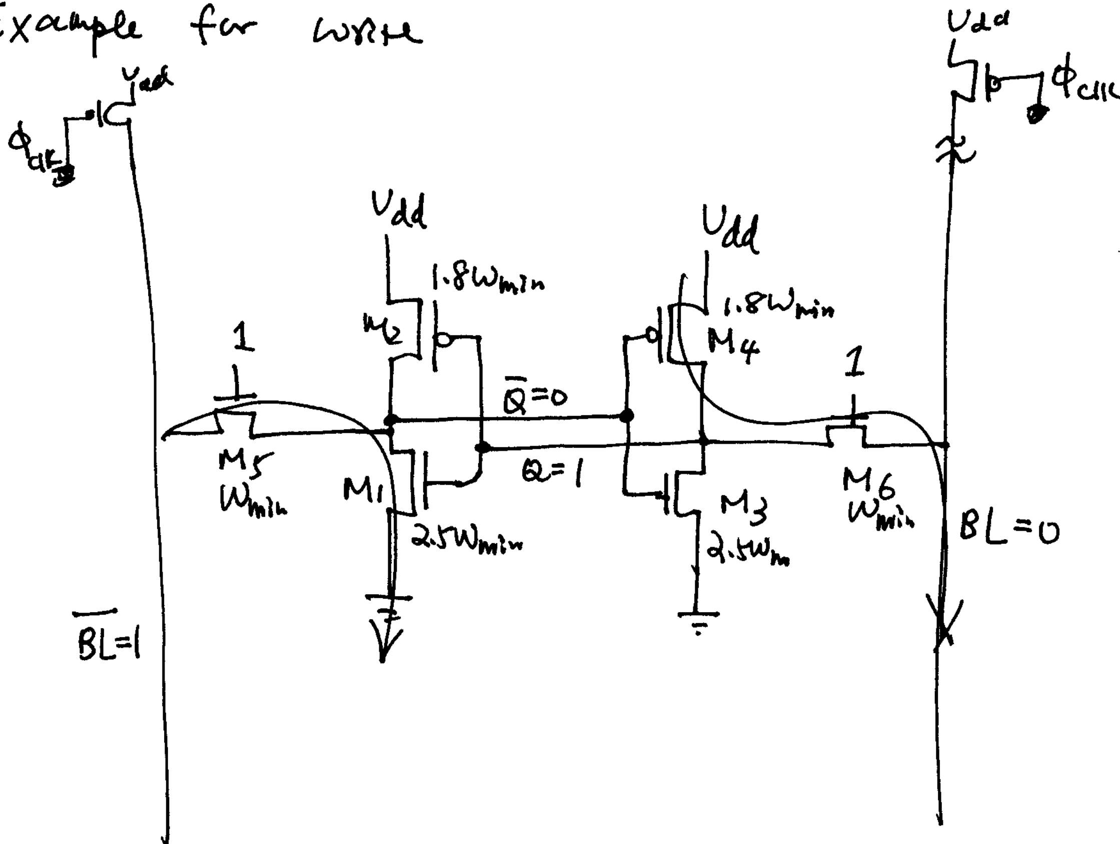
- A reasonable sizing is

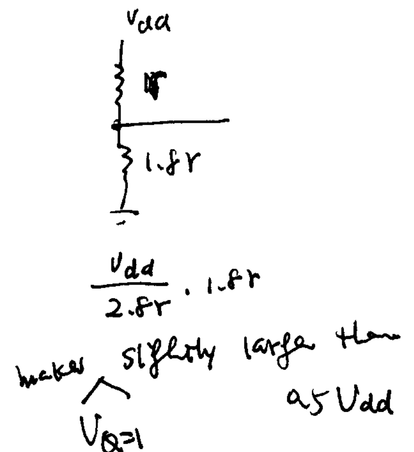
$W5 = W_{min}$, $W_1 = 2.5\,W_{min}$.

Similarly, $W6 = W_{min}$, $W_3 = 2.5\,W_{min}$

P on if $V_{gs} < V_{TP}$

- Example for write





$\frac{V_{dd}}{2.8r} \cdot 1.8r$

makes $V_{Q=1}$ slightly larger than $0.5\,V_{dd}$

- If write 1, no change!

- So, let us try to write 0.

- $\overline{BL}$ cannot change the state of $\overline{Q}$ as we discussed.

- $Q=1$ must be changed first by discharging through $M6 \rightsquigarrow BL$ line.

- $Q=1$ must drop below $V_{dd}+V_{TPO}$ such that M_2 can be on. So, there is a sizing problem between M_4 and M_6.

5
-0.7
4.3
③

- Once Q drops to below $V_{dd}+V_{TPO}$, M_2 begins to be on and charge $\overline{Q}$ to 1 and change the state of the cell. (Get stablized)!!

- M_4 must be ~~small~~ large enough, such that Q can drop to below $V_{dd}+V_{TPO}$.
small (in size)
i.e., its R must be large !!

$$\left(\frac{W}{L}\right)_4 \leq 1.8\left(\frac{W}{L}\right)_6$$

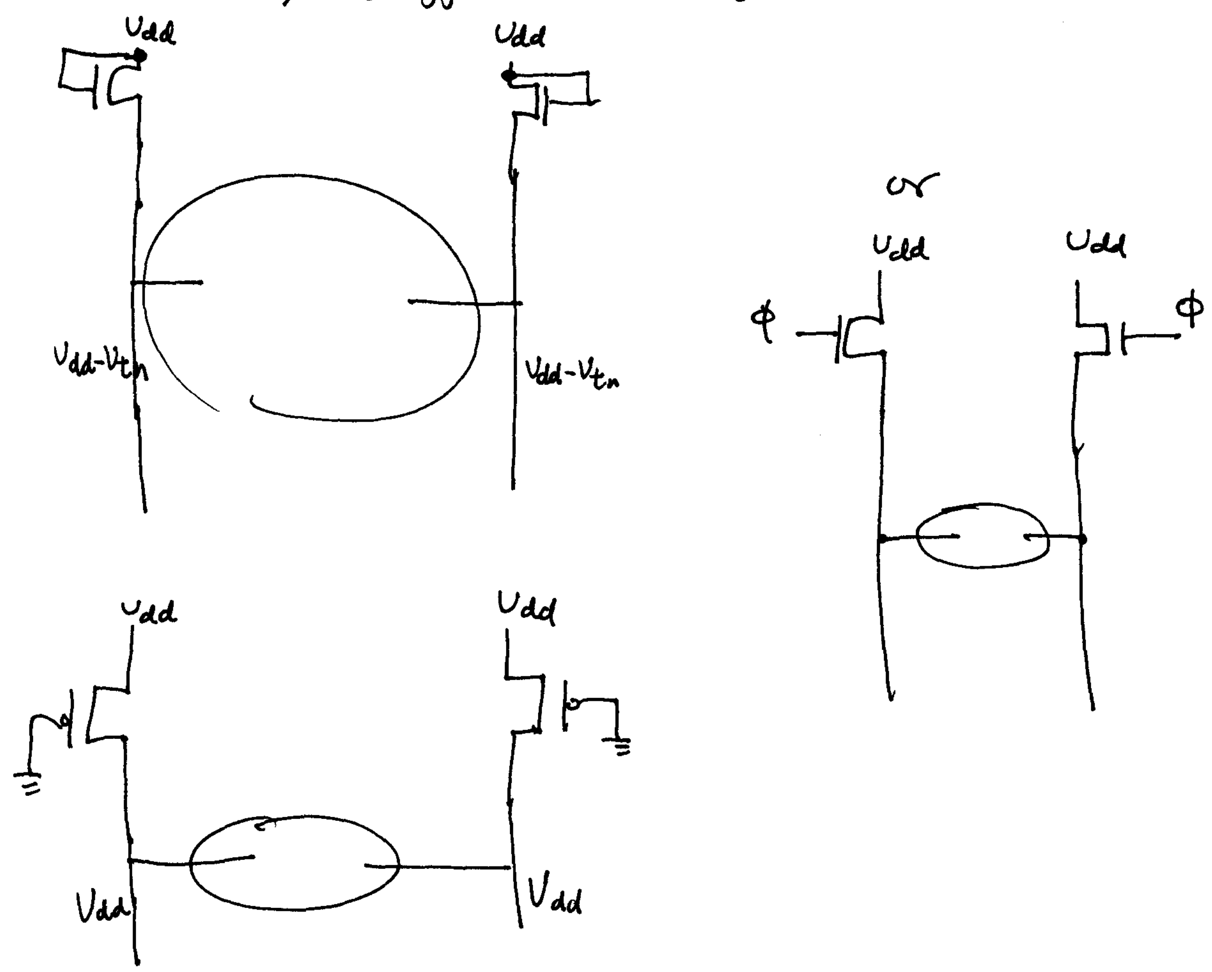
~~This makes [illegible]~~

$$\Rightarrow W_4 = W_2 = 1.8\,W_{min}$$

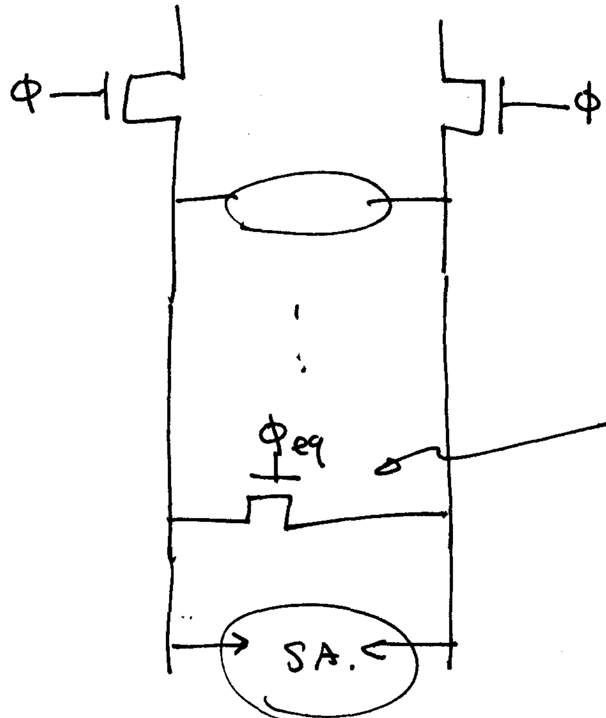
- Bit line Load.
  - In fact, the bit lines can be charged to lower than V_{dd} to save power
  - However, if BL (in P.184) is too low, $Q=1$ cannot be maintained, and the state may be changed.

  ∴ Bit line voltage ↑ ⇒ more stable memory state.

- can have many different designs.

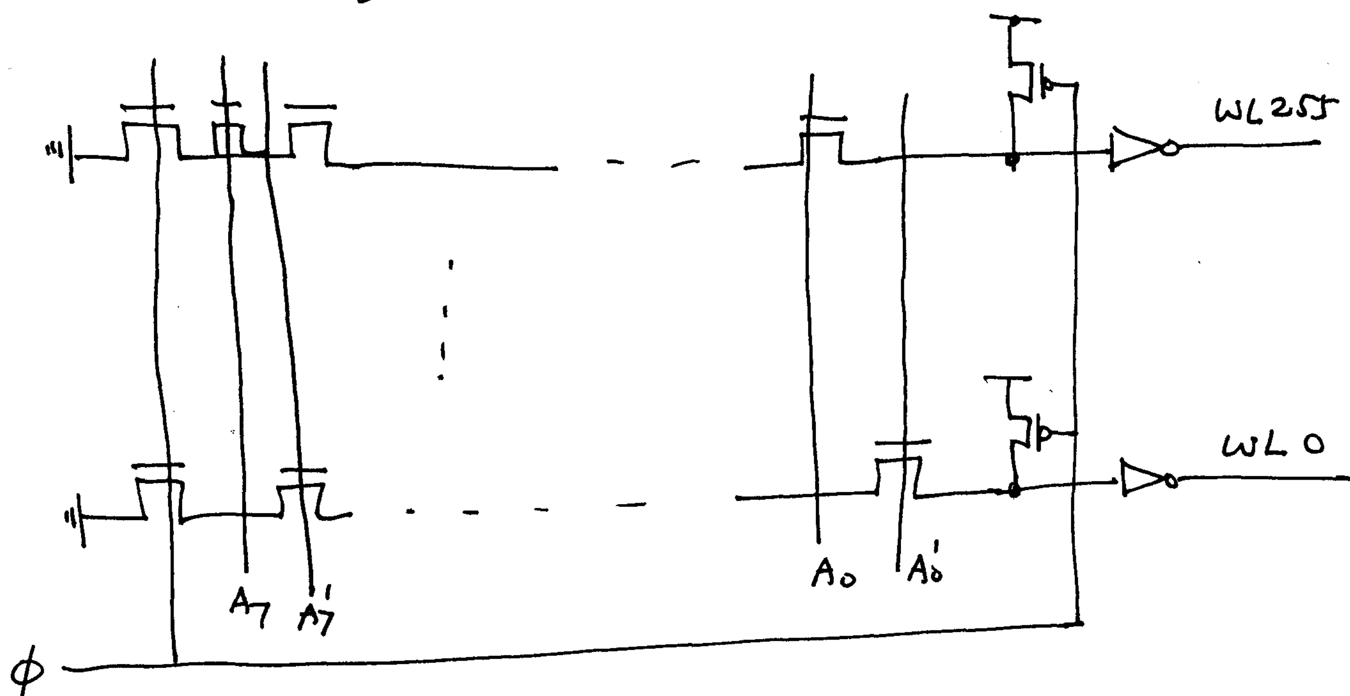


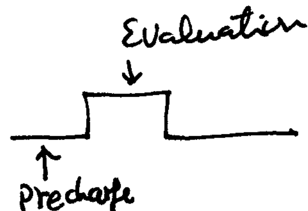
• No matter what kind of pull-up ckt, equalizing device must be added.



After each r/w operation, this must be activated for a while to equalize the voltage.

• Row Decoder   8 → 256 mapping (Example) — NAND-type





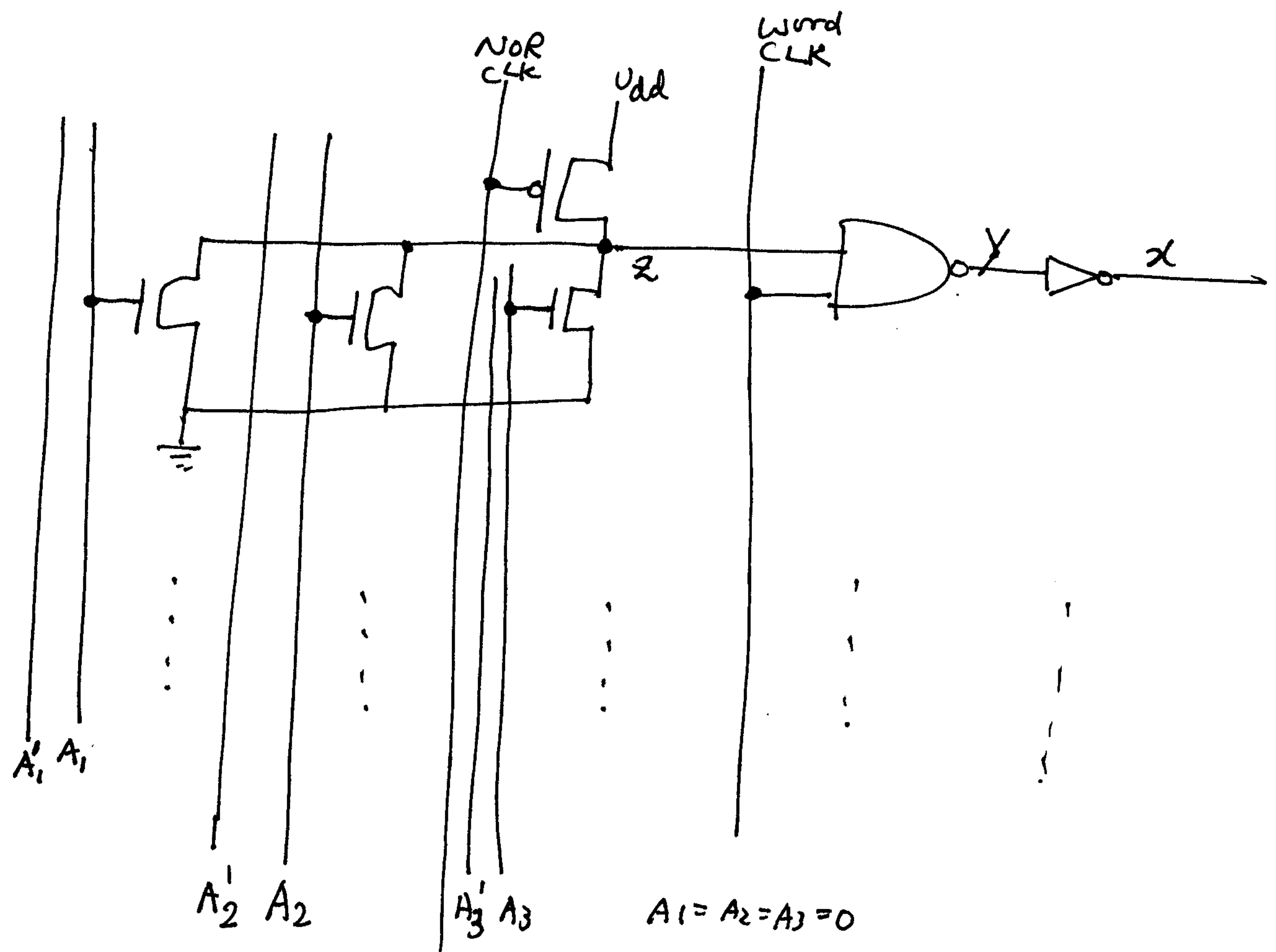
- Dynamic ckt
- $\phi=0$, all outputs are zero.

  $\phi=1$, only one word line would be 1

- very power-saving, slower (∵ larger height of evaluation NMOS stack)
  (only one line discharge)
- must add inverter buffer, since there is no power source to drive the high capacitive load.

- NOR-type

3 ⟶ 8 mapping



$A_1 = A_2 = A_3 = 0$

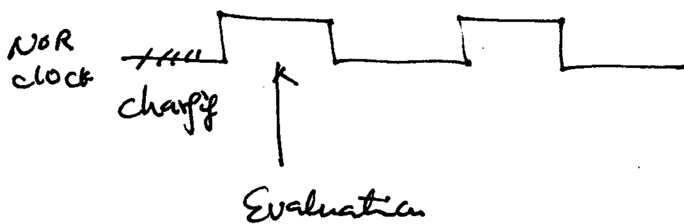
$\Rightarrow z = 1,$

$\Rightarrow y = 0$

$x = 1$

• faster, but consumes more power

∵ N-1 rows are discharging.



• Multi-stage Decoder (or called predecoder) (Compromise between NAND-type & NOR type)

NAND implementation cited in P.188

NAND implementation

=1 when $A_0 - A_n = 1 \cdots 1$

Group $2^m - 1$

only one stage be activated → save power!!!

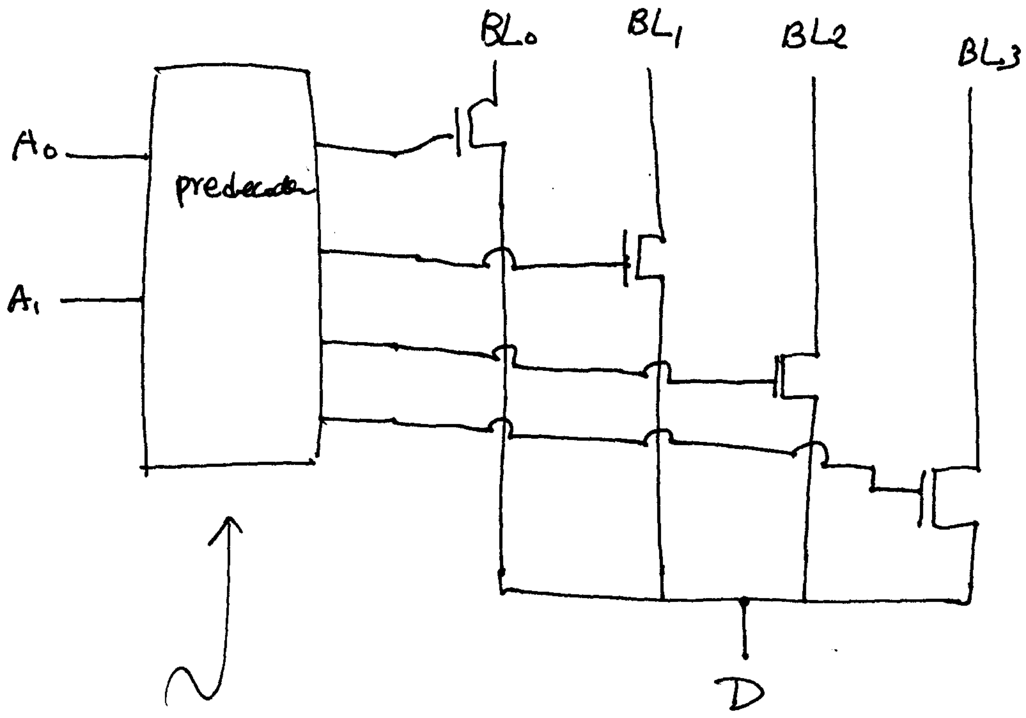
1 when $A_0 - A_n = 11 \cdots 10$

• Can be partitioned to more stages

Group 0

A_0 ... A_m ... A_{m+1} ... A_n

- ~~Row~~ Column Decoder : show by 2→4 mapping

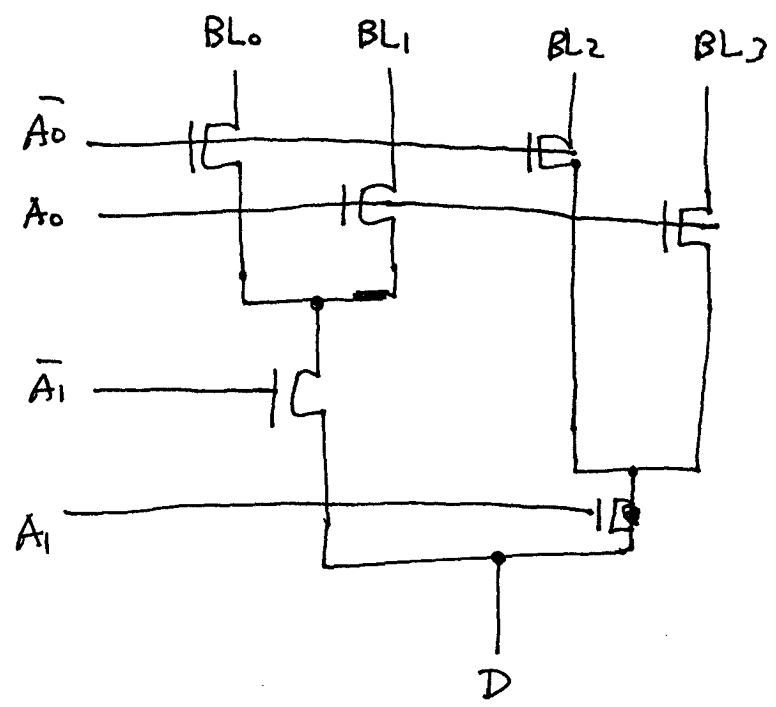


predecode + pass transistor

Don't need to be fast!

~~this is near the end of read operation.~~

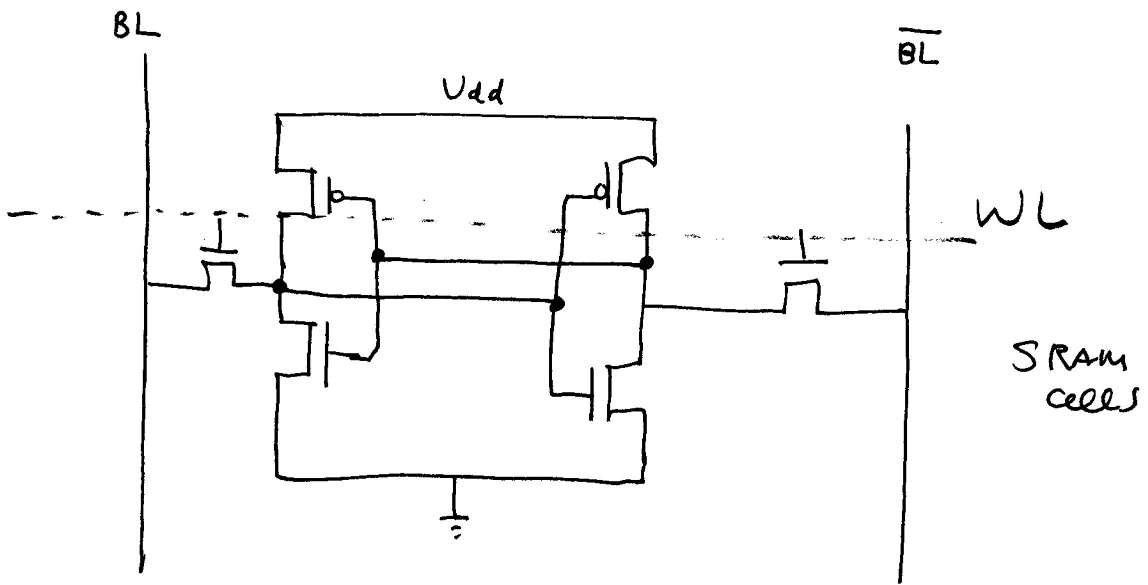
Tree decoder

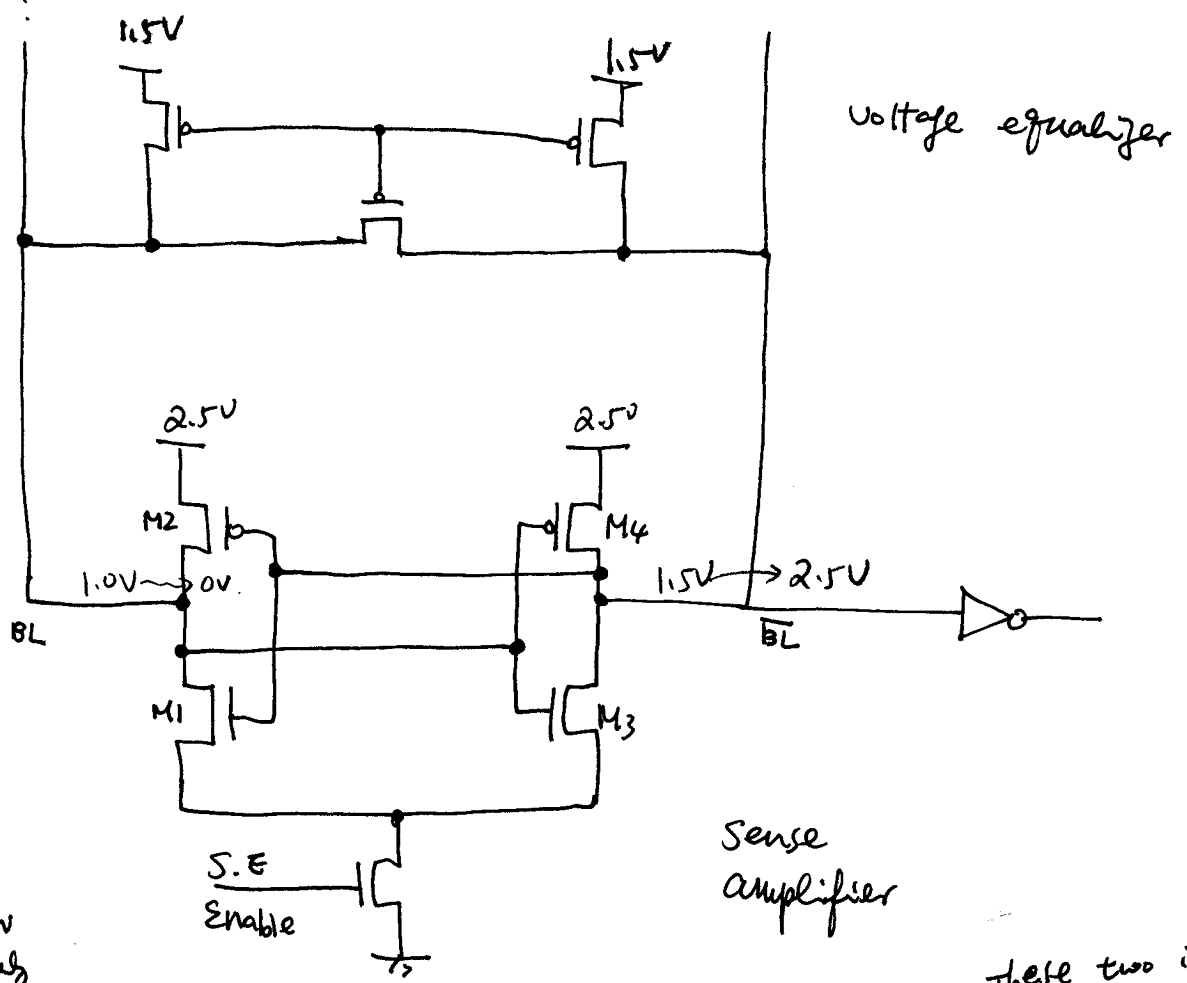


slow!

may need to pass many transistors

• Sense Amplifier





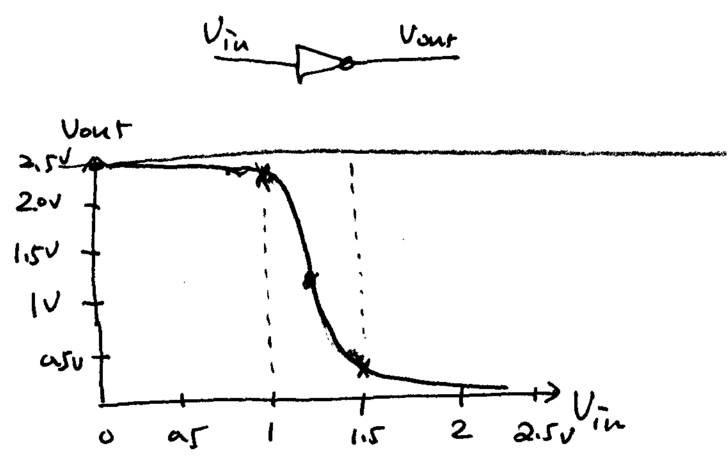
Step1:
BL changes 1.5→1.0V

step2:
$\overline{BL}$ changes 1.5→2.5V quickly

Step3: This quickly makes BL changes from 1.0V→0V. → step4: This feedback and quickly makes $\overline{BL}$ changes 1.5V→2.5V.

• These two inverters push each other !!!

- Inverter is a high-gain device



– Change in input 0.5V ⟹ output change about 2.0V

– ∴ 4 times of gain.

- Initially, both BL & $\overline{BL}$ are precharged to 1.5V.

- If BL is discharged to 1.0V, $\overline{BL}$ remains at 1.5V. ⟿ (increase to) 2.5V

  $\overline{BL}$ remains at 1.5V ⟶ BL dropped to 0.5V ⟿ (eventually to) 0V.

- If $\overline{BL}$ is discharged to 1.0V. ⟶ BL remains at 1.5V ⟿ (eventually) 2.5V

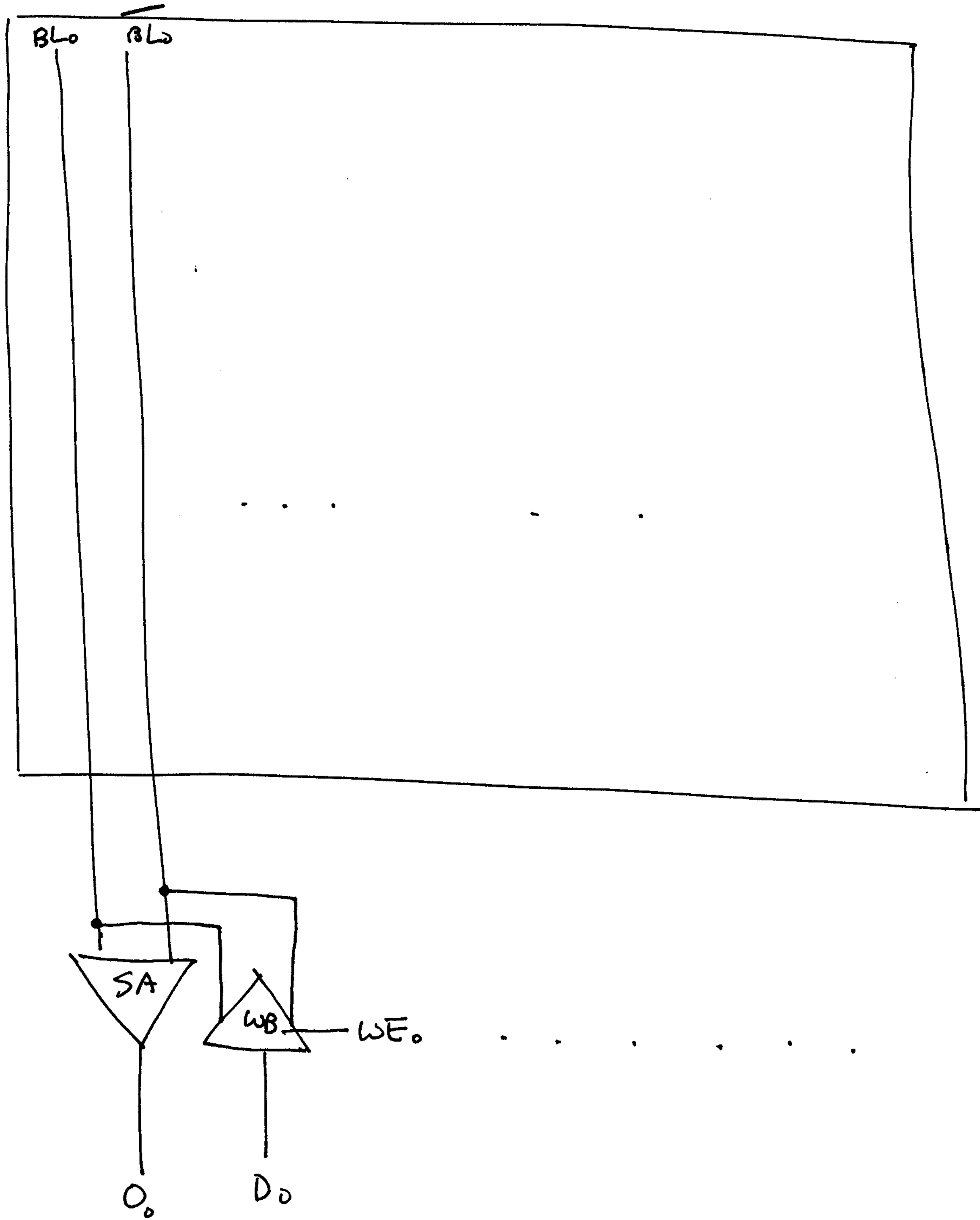
  BL remains at 1.5V ⟶ $\overline{BL}$ dropped to 0.5V ⟶ (eventually) 0V.

- This is most commonly used in real designs and Consume much less power than current mirror based S.A.

- The sense amplifier enable signal is generated using a self-timed approach to reduce the power consumption.

  that is BL will not charged to 2.5V, $\overline{BL}$ will not be discharged to 0V } or vice versa

• Read / write orperation



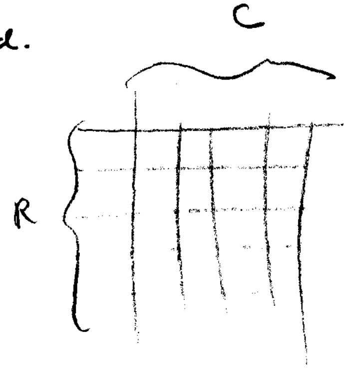
- Low-power memory design

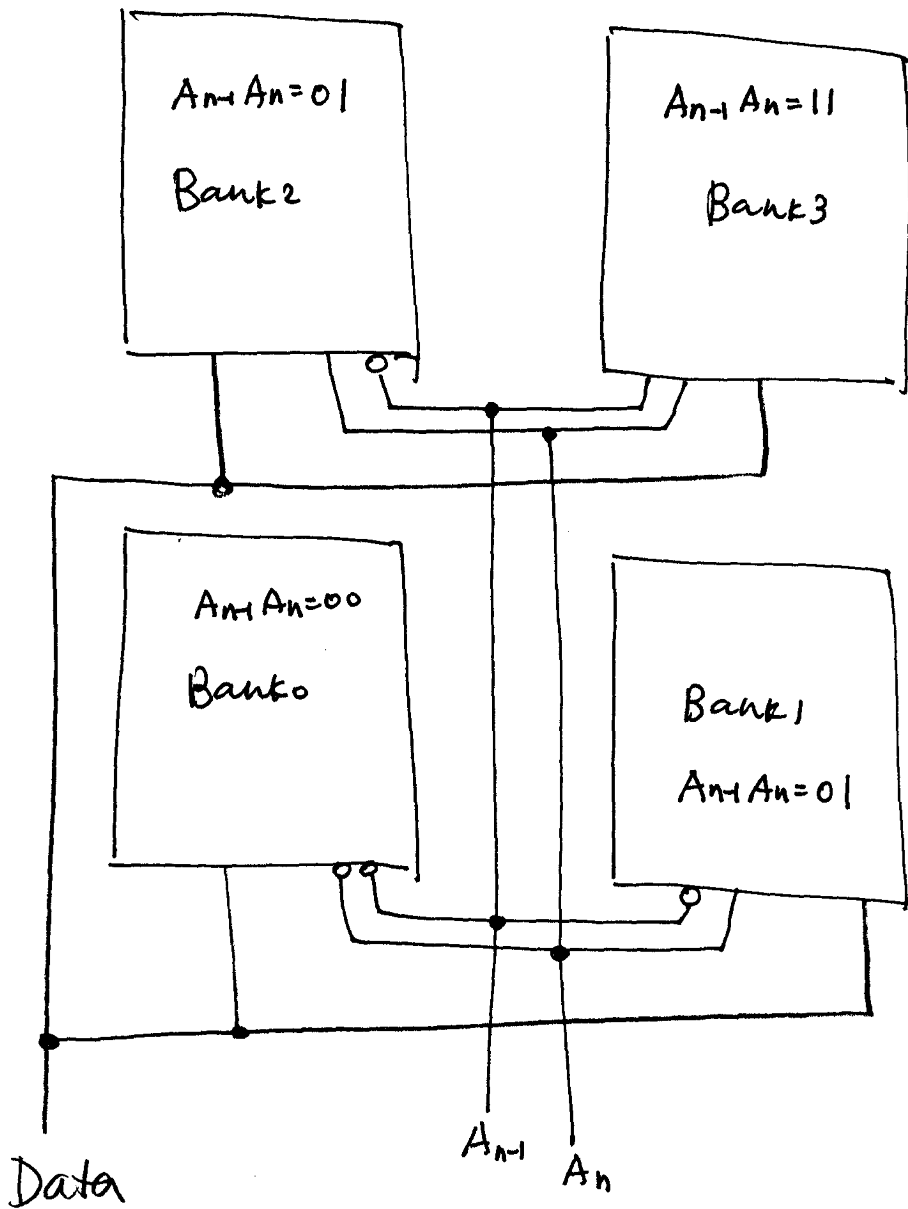
① Memory sub-banking

- R rows, C columns, C_{cell} : C (bit line switch/cell)

→ totally, $R \cdot C \cdot C_{cell}$ of cap. is switched.

- Try to partition the memory

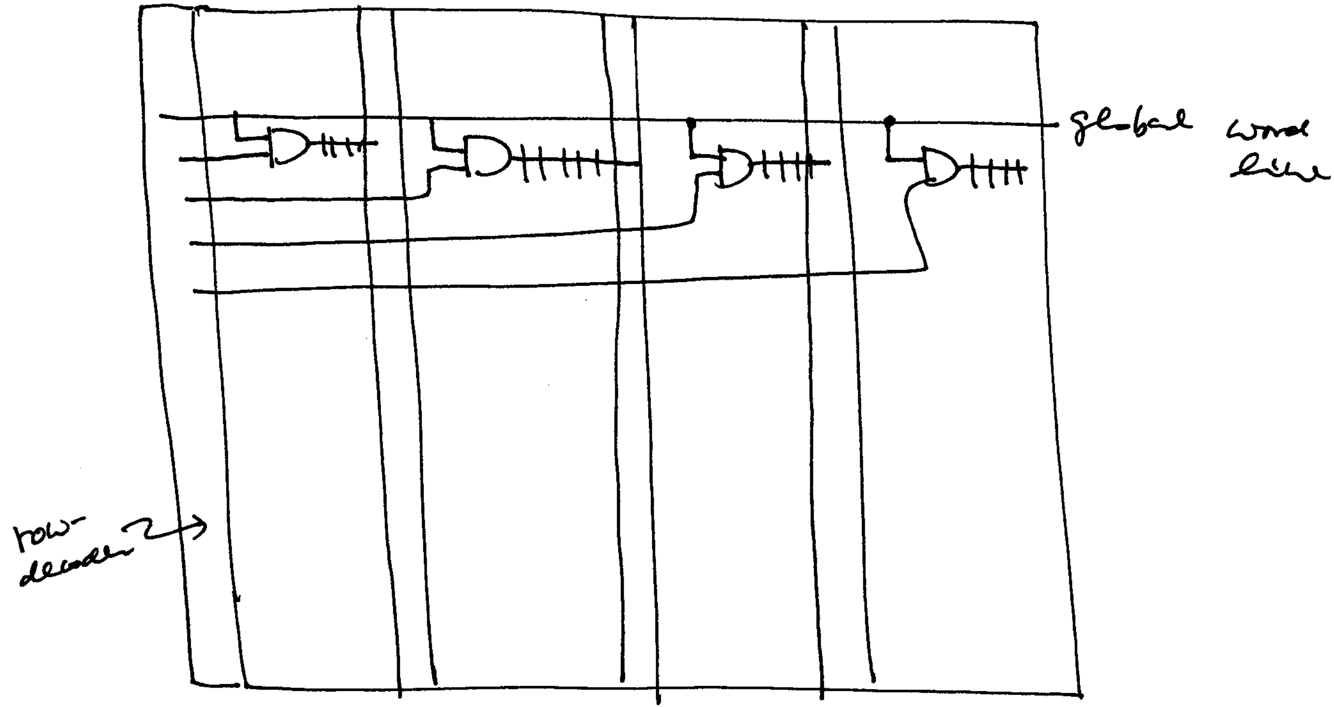




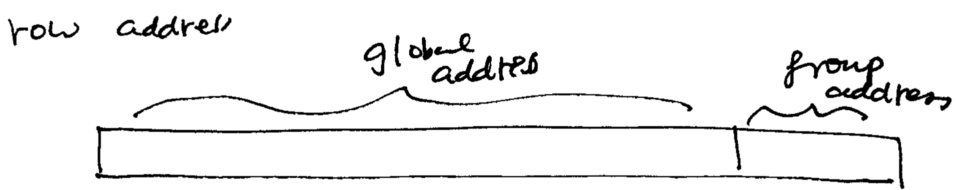
Make it B banks

$$\Rightarrow \frac{R \cdot C \cdot C_{cell}}{B}$$

- if interconnect is not considered.
- Can find an optimal solution.

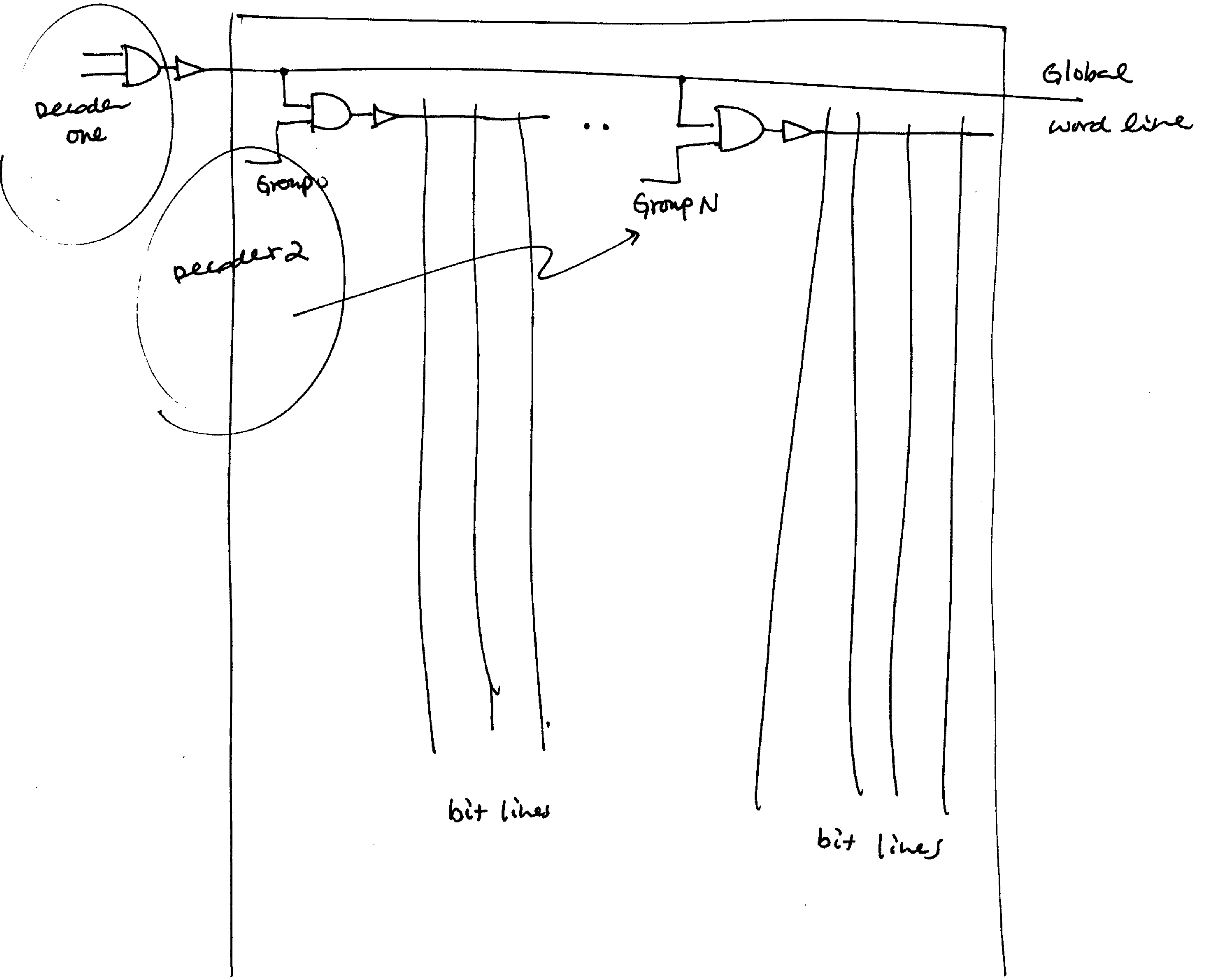


Divided word line



- Needs at least 2 metal lines
- local word line: polysilicon
- global word lines: metal 1
- bit lines: metal 2.

• Divided Word Line Architecture



• Reduce power consumption in Bit Lines.

• Bit line segmentation

precharge

sense amplifier

turn off

turn off

word line activation

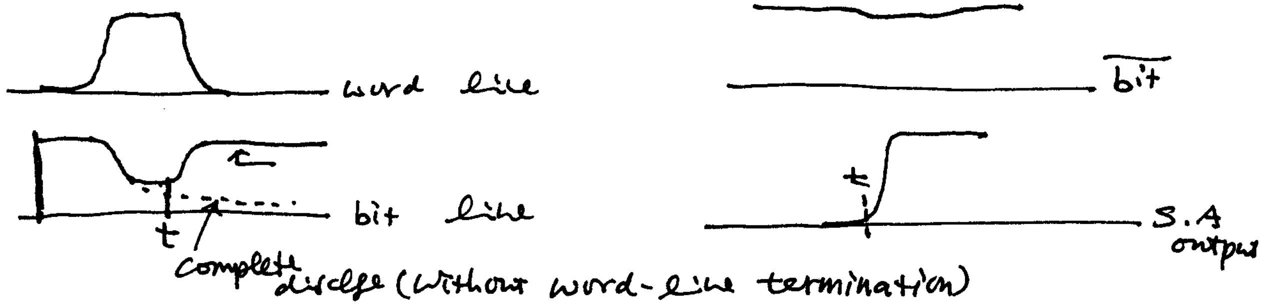
sense amplifier

Only either of these two segments to be discharged!!

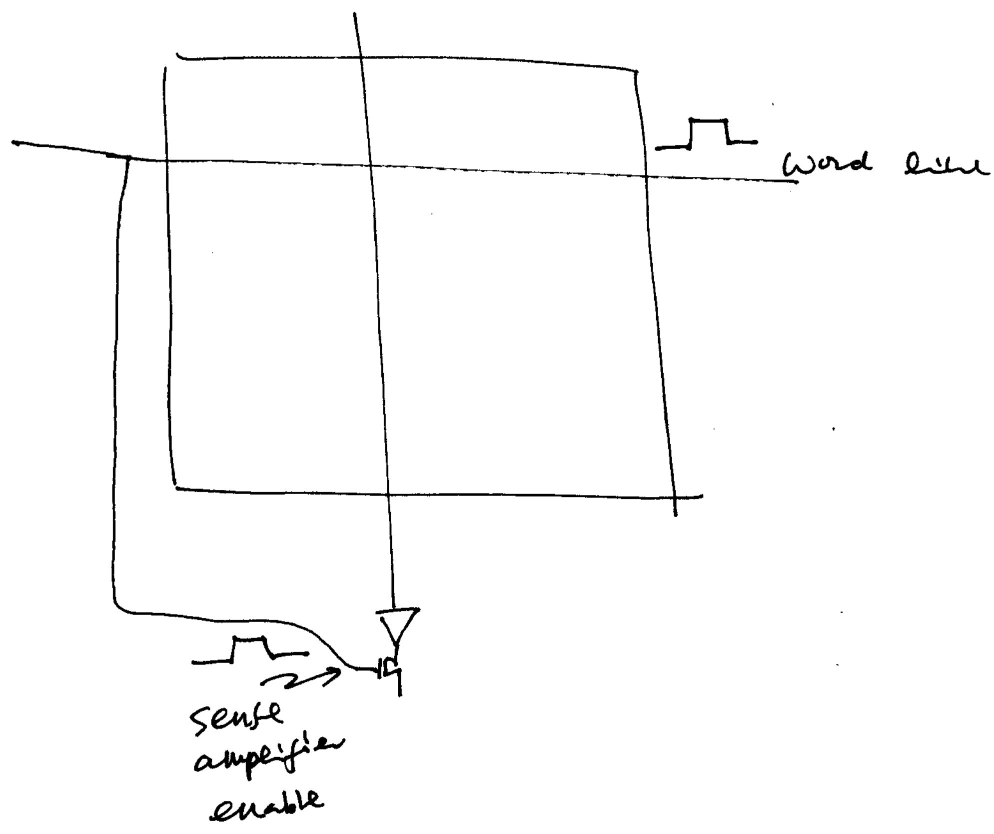
• Reduce voltage swing on Bit Lines

* Not quite feasible

∵ affect noise sensitivity, Complexity of sense amplifier, performance of the RAM Core.

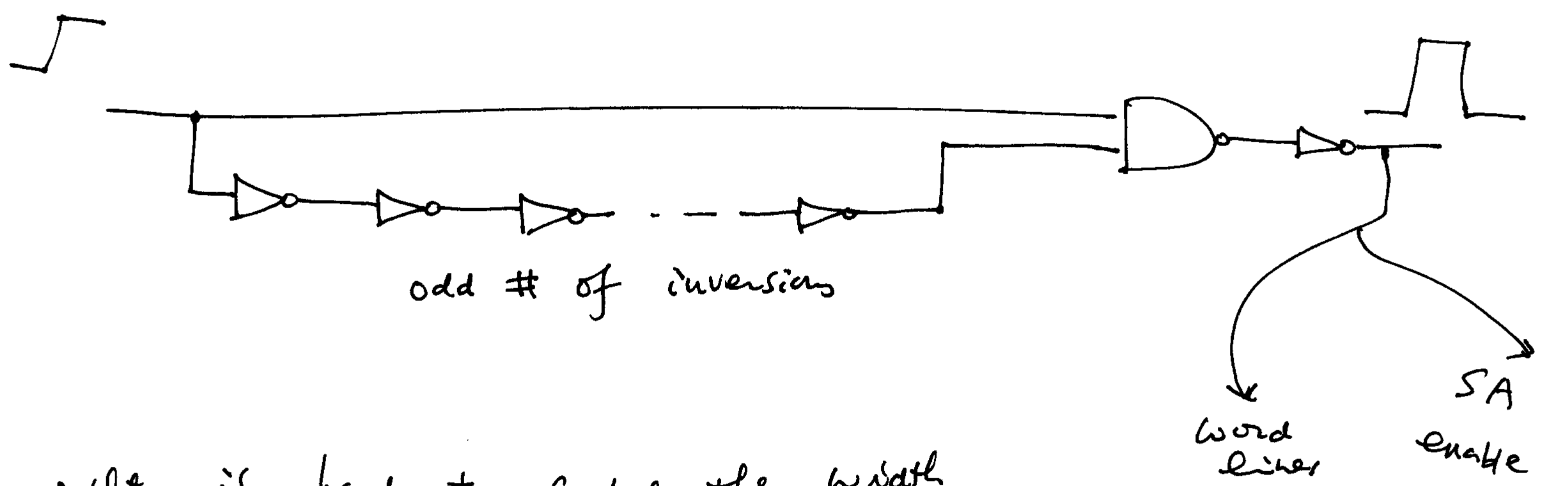


• Pulsed Word Lines



• Try to limit the bit line voltage discharge.

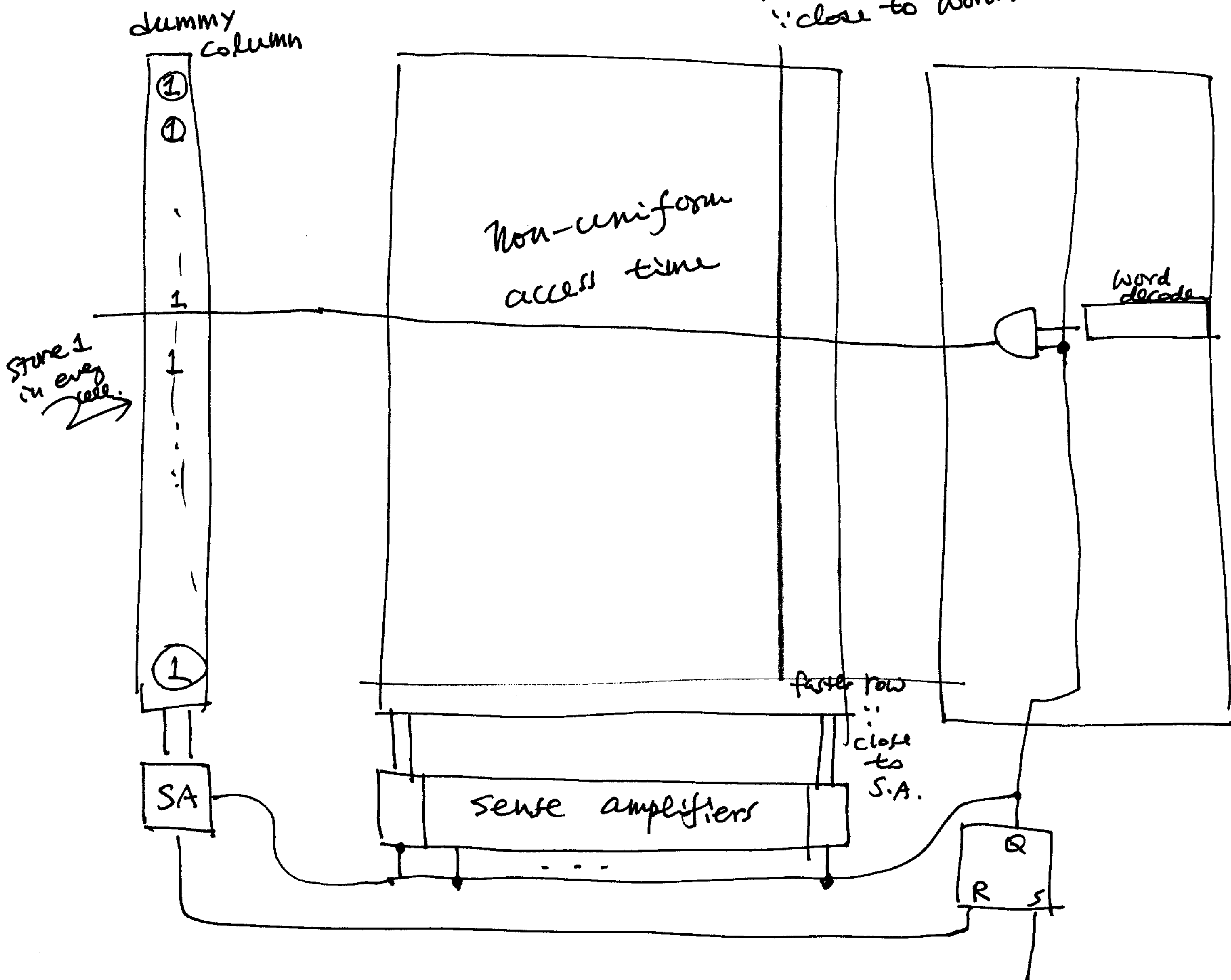
• Try to enable the word lines for precisely the time needed to develop the bit cell voltage discharge.



• It is hard to control the width

• Try to use self-timed word lines.

- Self-timed word lines.



① SR flip-flop is set, and the word line (SA) is triggered.

② When the dummy SA (slowest) generate 1, the rest of the columns would have been sensed.

③ The high signal reset the flip flop and turn off the word line (SA).

Summary: power optimization for SRAM

Reduce total capacitance switched

① Banked organization

② word-line division

③ Multi-stage decoder

④ Bit line segmentation

Reduce the voltage swing across switched capacitances

① Self-timing RAM

turn off word line & SA enable, s.t. the voltage swing in the bit lines can be minimized.

② Reduce bit line voltage swing.