6.978

LECTURE #17. NOVEMBER 16.

TODAY: MEMORY CELLS & SUBSYSTEMS

PROJECT LAB: A NUMBER OF ST. WANT TO USE LAB THIS WEEKEND. I'LL BE IN BOTH SATURDAY & SUNDAY FROM ~10+ TO ~4+.

- ALSO, OFFICE HOURS TMRW FROM ~11 TO ~4.
- HOW MANY MIGHT BE INT. IN USING LAB OVER THANSGIVING? IF A FEW WHO CAN COOP. ILL GIVE KEY TO ONE. MAIN THING: CAN'T LEAVE OPEN & UNATTENDED DURING OFF-HOURS.

[ALSO: THE LAB ASSISTANTS COULD STAY LATER (SAY ~9) IF SOMEONE COULD REMAIN IN LAB ~6 TO 7 SO THEY CAN GET DINNER.]

MEMORY CELLS & SUBSYSTEMS

- These presently get lots of attn because of the huge market and money being made manufacturing memory chips.

- Presently, general purpose computing done by Von-Neumen type machines, where memory is made distinct from processing in the hardware [CPU] ⟷ [MEM].

  So much effort is made to increase speed (reduce T_p) of CPUs, and to increase size and reduce T_m (cycle time for mem) of memory. There is an insatiable demand for denser, faster, cheaper memory.

- Some Things To Keep in Mind During Today's lecture:

  (a) Present extreme emphasis on memory chip design is result of present computational environment. This may change in the next decade.

  (b) As the interior device sizes within CPU & MEM are decreased, a higher & higher proportion of the energy & time cost per computation is due to the energy & time to transport info from MEM ⟷ PROC.

(c) Because of the #'s of best LSI designers preoccupied with Memory cell design, it is a CLASSICAL ENGINEERING ART.

All these people are working on variations of just a few cell types, and the alternatives have been explored thoroughly.

Since whole chip is memory, the process, circuit, and subsystem designs have been simultaneously optimized. Even without much computer aids, this is possible because the cells are simple, and subsystems very regular.

(d) To optimize designs for min delay & pwr dissipation will require the use of electrical simulation.

(e) The ultimate 1-T memory cell which I'll show requires very clever (experienced) circuit design in its interface and support circuitry.

- So this is a diff. design environment than that discussed in this course. Rather than doing lots of designs per unit time, designing hierarchically to build big systems, and using design methodology constraints to keep out of trouble, MEMORY DESIGN involves highly optimized lower level process and circuit design.

- Note also: Changes in way we might do computing will be discussed during week of Dec 4-8:
  DEC 5 Carver Mead (H.Com.Proc.) Dec 7 Wayne Wilner (Rec. machines)
  DEC 8 Carlo Sequin, U.C.B., X-TREE [How many can come?]

- To Prepare for that week: Read CH8 p1-9, p31-32, p57; Skim p33-56. (notice res-ltc)

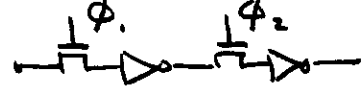
  Also: Make a copy of, and at least skim through the recent highly important paper by John Backus of IBM - "Can Programming Be Liberated from the von Neuman Style? A Functional Style and Its Algebra of Programs"
  CACM AUG '78 V21 #8

- WE'VE USED SEVERAL WAYS SO FAR TO STORE INFO:

  EX: • The Shift Register, to make serial memory
  • The DM Register cell, to make randomly accessed (by word) memory array.

- LET'S EXAMINE THE DENSITY, PWR OF THESE and see how they compare to available memory chips.

- THE SHIFT REGISTER: Recall Fig 8b, CH.4 for layout:

  one bit of storage requires two SR cells: 

  Area $\approx (21\lambda \cdot 2) \cdot 19\lambda =$ 126 μm x 57 μm

  '78 λ = 3 μm use all 3 μm lect. th.

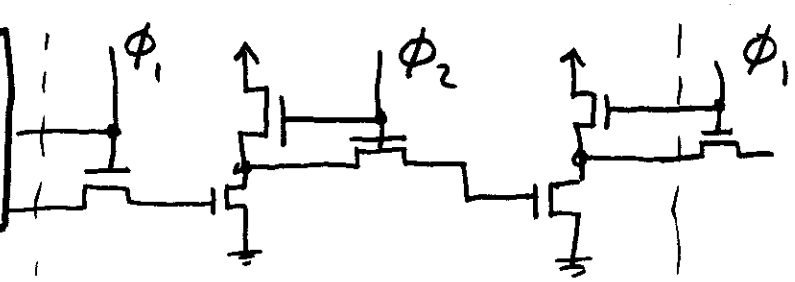
- How many Bits could we place in a big chip, say 4 mm x 4 mm? also use for core + Reg

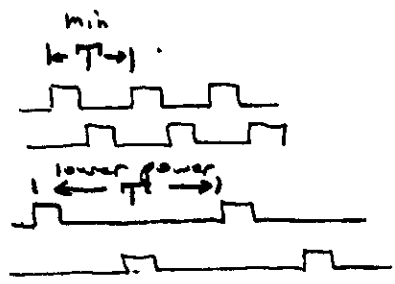
  # BITS $\approx \frac{4000^2}{126 \cdot 57} \cong$ 2K bits

- Remember we previously calculated a power dissipation of ~ 8 to 12 w/cm². This was rather conservative (high) and also could be reduced by using longer pullups, narrower pulldowns to perhaps 3 to 6 w/cm². Still a bit hot.

- There's really no way to reduce cell area much. But there is a way to reduce pwr: trading pwr red. against increased delay: USE ENH MODE PULLUPS (may req longer ratio than 8:1), and CLOCK the PULLUPS:

6 Transistors per BIT of memory



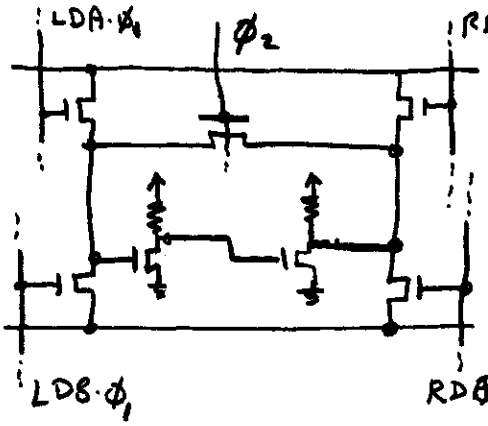


NO STATIC PWR DISS WHEN CLOCKS OFF! IF LOW DUTY CYCLE, THEN LOW POWER (BUT LONG T compared to MIN T)

(SEE THE TI BOOK FOR MORE SR TRICKS!)

- So the SR isn't really very dense. It is, however, extremely easy to interface as a subsystem for storing small amounts of info.
- How do we get more density, especially without incr pwr/area substantially? We use RAM:
- LETS BEGIN WITH THE OM REG. CELL, Then progress thru series of denser RAM cell designs
- OM REGISTER CELL:

IF 2-BUSSES:	9-T's /BIT
IF 1-BUS	7-T's /BIT



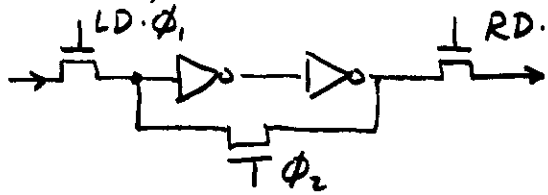
Busses Precharged. Reading just has to pull down. So fast.

NOT STATIC

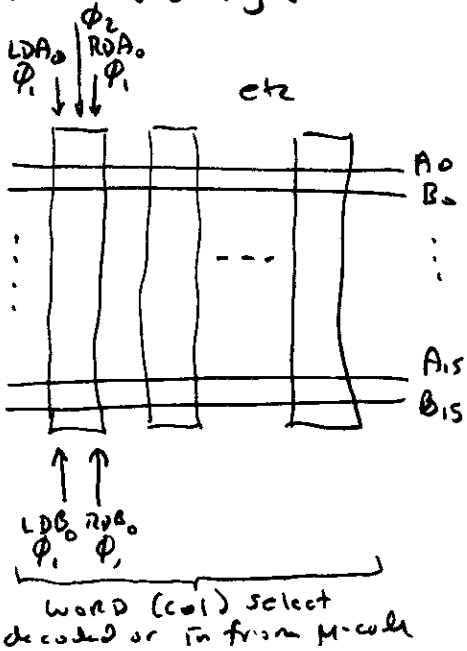
Basically very simple. Really like SR which feeds back on itself:

Loading into C_G only

Reading from Inv output only



So Very easy to make a memory subsystem which is directly interfaceable with, compatible with our 2-Φ clock scheme: No clocking tricks needed: For example the 16-16bit 2 port REGISTER SUBSYSTEM IN OM-2



DENSITY: CELL IS ~ 54λ x 48λ

∴ In 4x4mm: $\frac{(4000)^2}{162 \cdot 144} \approx$ 700 BITS

EVEN LESS THAN SR

Even if used one Bus (7-T's/cell) could only get ~ 1K in 4x4mm.

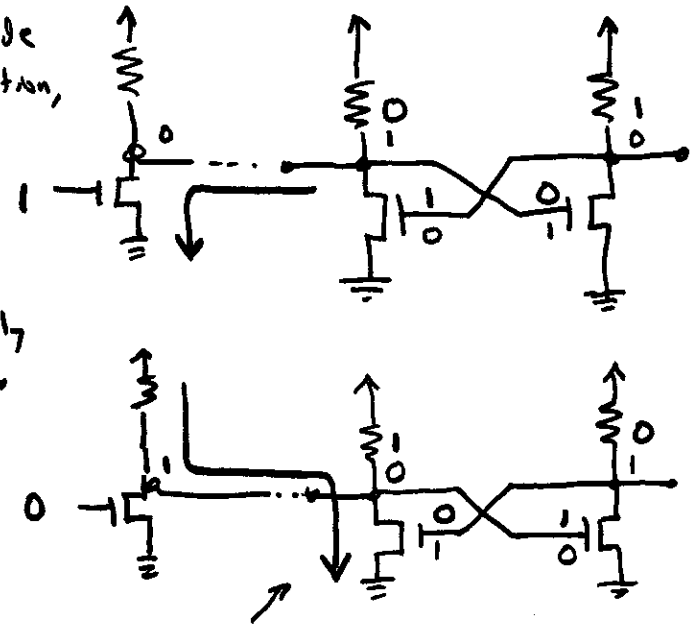
NO Problem with PWR.
BUT MUST BE CLOCKED

HOW TO MAKE A STATIC RAM

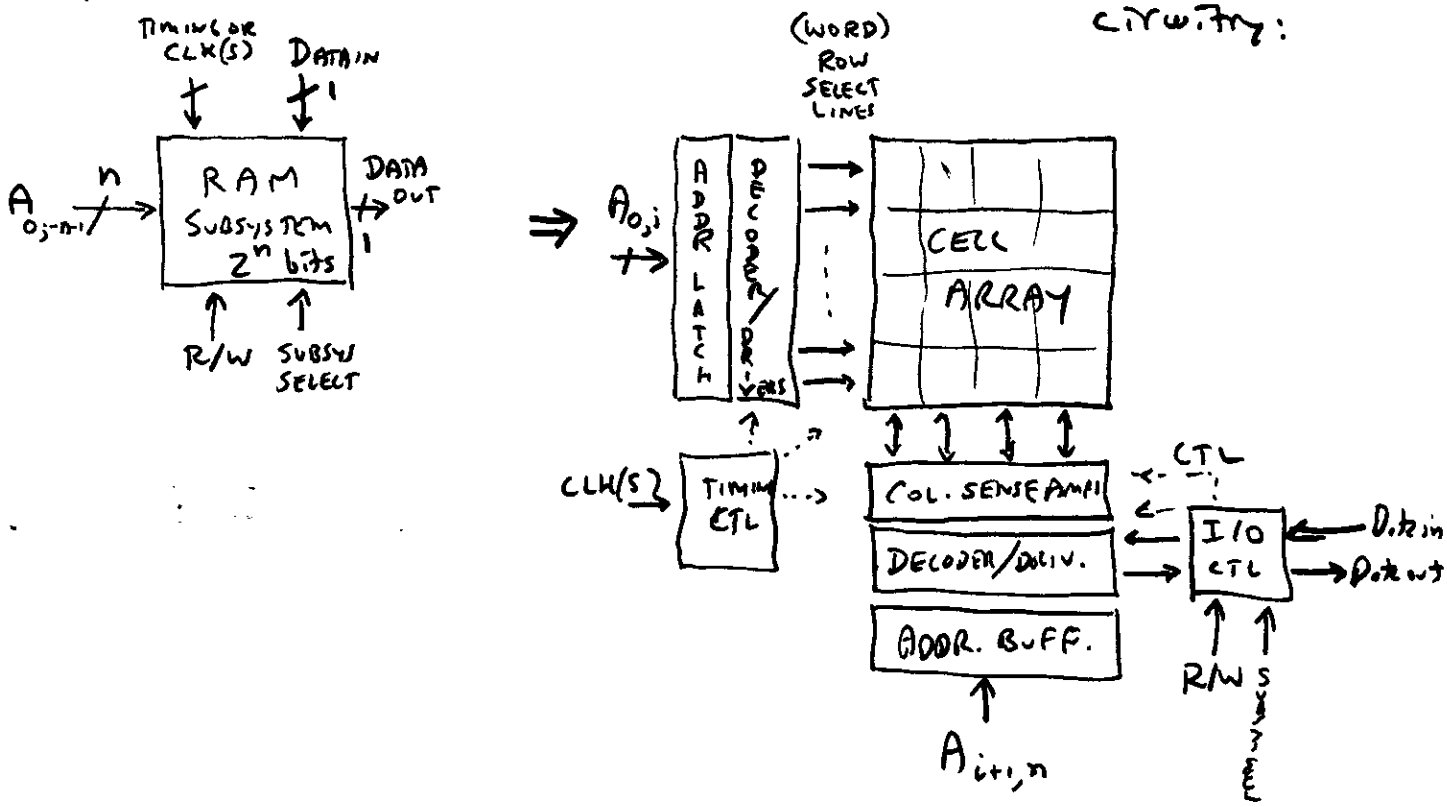
THE SIMPLEST STATIC RAM CELL IS BASED ON CROSS-COUPLED INVERTERS:

Same as our Reg cell, but no clocked pass-T in feedback path:

- If either input/output node is pulled down, by ext. action, other side turns off, latching the pulled down side ON.
- Will hold state indefinitely (STATIC) unless power goes off.
- Note: Ratios have to be right: external driver has to have width pull-up in order to source enough current to raise input node above threshold. Usually use double rail input



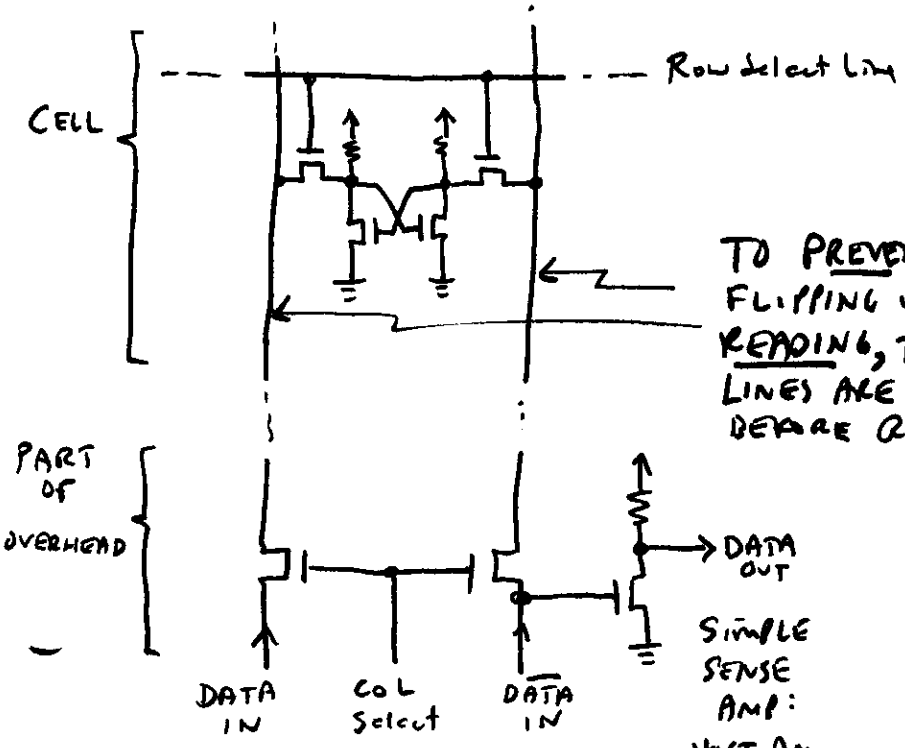
IN GENERAL: MAKING RAM SUBSYSTEMS: There is overhead circuitry:



VERY IMPORTANT POINT ABOUT OVERHEAD CIRCUITRY:

- There is some fixed overhead for given type of memory cell
- For Address size n, cell array area goes as n^2 but overhead usually goes as n ... to n ln n.
- So, "bigger" memories require less fractional area in overhead.
- HOWEVER: The trickier, dynamic memories which have increasing density, smaller cell sizes, require greater fixed overhead.
  - > no such thing as a free lunch. i.e: can't get small memory subsystems with ones/bit on order of the 16K RAMs, because of fixed overhead.
  - > So, for small RAMs use REG CELLS, OR AT LEAST STATIC RAMS - Easier to interface.

THE 6-T STATIC RAM:



Area: Maybe 125 μm x 125 μm

So: $\frac{(4000)^2}{125 \cdot 125} = (32)^2 = 1K$

per 4mm x 4mm

Not counting OVERHEAD

TO PREVENT FLIPPING WHILE READING, THESE LINES ARE PRECHARGED BEFORE READING

i.e. ~ same as single bus REG CELL.

If on-chip subsystem, could READ/WRITE WHOLE WORD AT ONCE

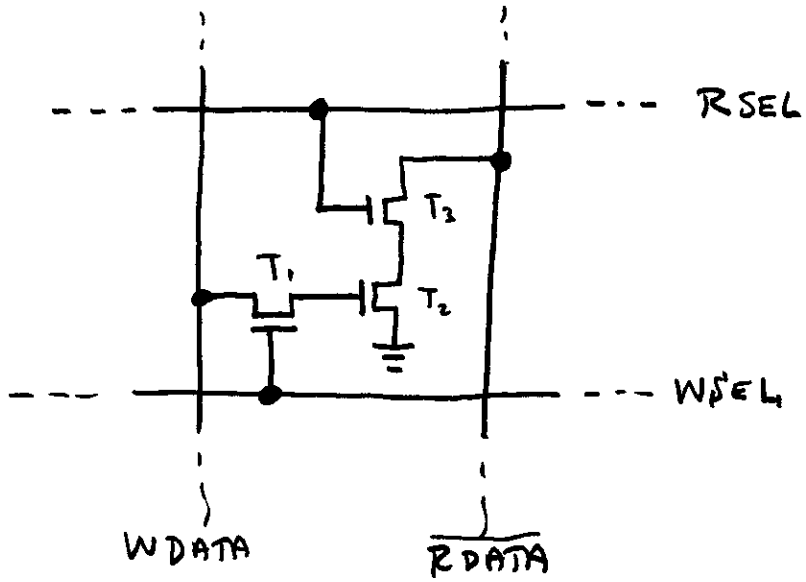
NO PWR/AREA PROBLEMS SINCE NOT DENSE ENOUGH! (IN BITS/AREA)

- **HOW DO WE MAKE DENSE RAMS?**

WE USE FORMS OF DYNAMIC RAM WITH FEWER T's AND WIRES PER CELL (NOT REALLY #T's That count but # WIRES, ALTHOUGH # WIRES ROUGHLY ∝ #T's)

- **The 3-T Dynamic RAM Cell:**

We've used charge stored on gates in our SR's, and in our clocked inverter REG cells: Let's make a RAM cell that uses this effect more directly:

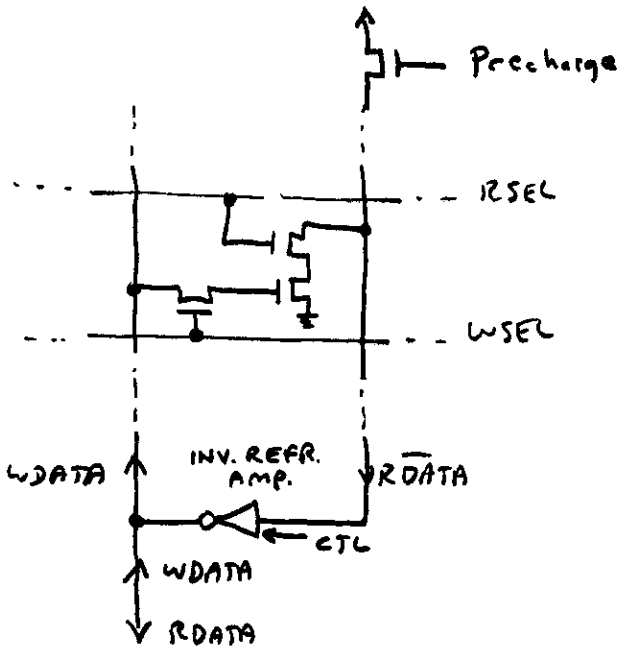


- 2-Select Lines, 2-Data lines per cell.
- Data stored on gate of T_2
- TO WRITE: Put data in onto WDATA. Raise WSEL for a while, then lower.

- **TO READ:** > RDATA is Precharged high.
  - > RSEL is then turned on. The RDATA line is then discharged only if input to T_2 is high.
  - > Note: Output is complement of input
  - > Note: The Read is non-destructive
  - > Note: The cell must be periodically refreshed.

3-T RAM (cont.)

- Lets Add some of the cell's overhead circuitry (see TI p. 125)



A Possibility:

- Read Out data for entire ROW Selecting appropriate column to route to output.
- REQUIRES ONE REFRESH AMP PER COLUMN.

- REFRESH BY RSEL, WSEL ON GIVEN ROW, TURNING ON ALL REFRESH AMPS AT ONCE. MUST DO THIS FOR EACH ROW EVERY FEW MILLISECONDS, REQ ADDIT. EXT. CTL.
- TO WRITE: SIMPLY SELECT COLUMN, PUT DATA-IN ON WDATA, AND SELECT ROW'S WSEL.
- THERE ARE MANY VARIANTS ON THE 3-T DYN RAM. SEE TI BOOK. FOR EACH CIRCUIT VARIANT THERE ARE MANY STICK'S, FINALLY MANY LAYOUT ALTERNATIVES. (NOTE ADJ COL. OR ROWS CAN SHARE A GND RETURN)
- DENSITY:

Conservatively: 60 μm × 60 μm per cell. So, # in 4mm × 4mm (not counting overhead)

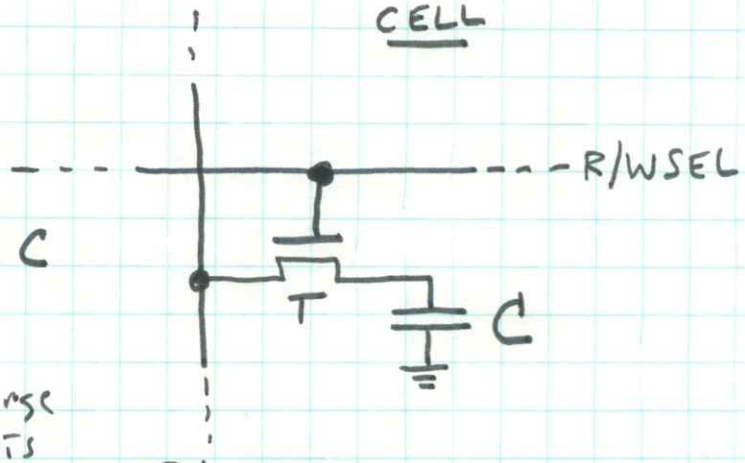
is: $\frac{(4000)^2}{(60)^2} \approx$ **4K bits**

- PWR: No problem: Switching power, and short dissipation thru T_2 -- T_3 of energy stored on RDATA.

4 Times as Dense as STATIC RAM. ≈ SAME PWR/AREA.

SUPPOSE THIS IS STILL NOT DENSE ENOUGH!

THE ULTIMATE (AT LEAST NOW) IS THE SO-CALLED
1-T DYNAMIC RAM



- Really two Poly over Diff Regions: one to form a switch, and a larger one to form a Capacitor C
- Works by simply storing charge (or lack of) on C when T is on. Later, sense charge on C by closing T again and seeing what happens to Data line.
- **Major Problem:** If R/W DATA line (and sense amp inputs) have capacitance C_L, then when close T to read, the voltage on C divides:

  Ex: If C_L low, C holds VDD, then a 1 on $C_L \approx VDD\left(\frac{C}{C+C_L}\right)$

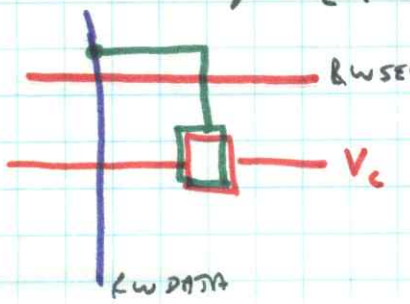
- In Real RAMs $C_L >> C$. Maybe by x20, x50, x100! So this gets really hairy!

> Need special form of sense amplifier circuitry.

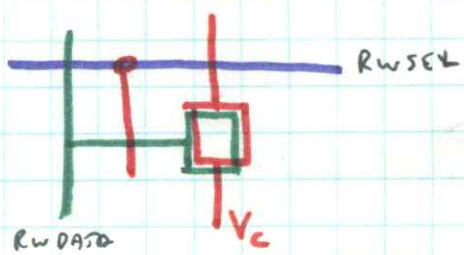
> Also, Readout is DESTRUCTIVE, & MUST Rewrite after every read, not just to refresh.

- **LAYOUTS**: Tricky: I think that to eliminate a Poly to Red contact, the Capacitor input is on Green, & other side goes out on Red to some voltage $V_c \neq GND$. 2 Possibilities:

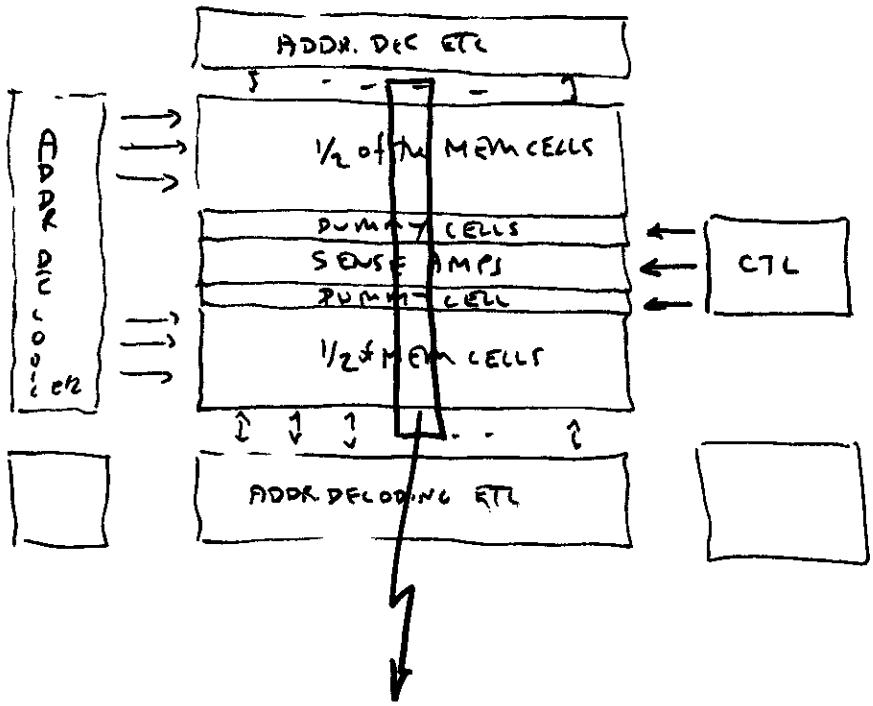
(a)



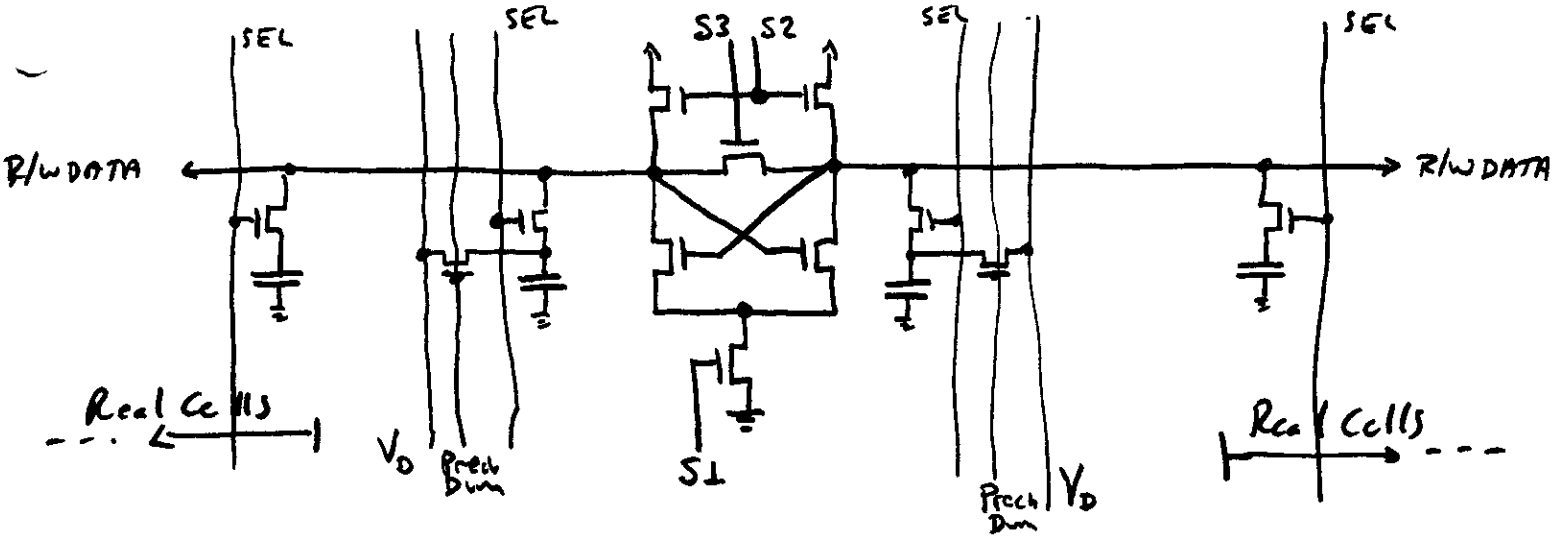
(b)



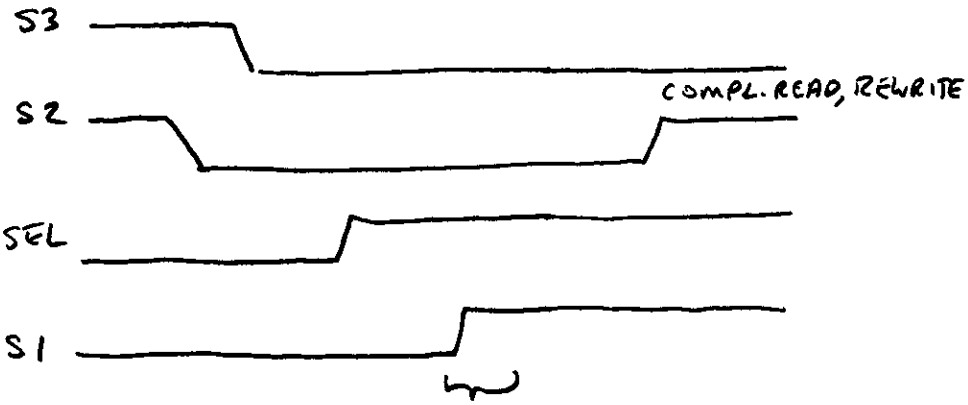
POSSIBLE WAY TO SENCE / REWRITE: ORGANIZE SUBSYSTEM FURTHER AS FOLLOWS:



TILT OVER ONE COLUMN OF THIS ARRAY:



- Dummy Cells Precharged to some V_D then to READ/REWRITE:



- when S2, S3 on, S1 OFF everything prech. high
- whichever s.th higher on SEL, other s.th discharges faster!
- when S2 comes back on, value is locked-in and can be rewritten

- SO, VERY TRICKY DESIGN. TRY TO MAKE C BIG (BUT CONFLICTS WITH SMALL CELL SIZE), MAKE C_L SMALL, MAKE SENSE AMPS THAT WORK WHEN $C_L >> C$.

  VERY HIGH OVERHEAD DESIGN COMPLEXITY. BUT LOOK AT DENSITY: CONSERVATIVELY: '78 $\lambda = 3\,\mu m$

- CELL SIZE $\approx$ 30 μm x 30 μm, SO IN 4mm x 4mm:

  $\frac{(4000)^2}{(30)^2} \approx$ **16 K bits** not counting overhead.

- Actual 16K RAMS MADE WITH λ ON ORDER OF ~ 2.5μm (5 μm wires). The new 64K rams will be similar and use λ ~1.25μm or a little smaller, (i.e. ~ 2.5 μm wires).

IF TIME: INTRO SOME MATL IN CH 8:

- Delays caused by huge relative capacitive loads. Applying the theory we developed in CH 1, we might think of organizing our memories hierarchically, rather than just making "longer wires" (SLIDE 3)

- If do this, we may pay a penalty / bit in that area / bit increases as the Branching Ratio α decreases (i.e. we branch more often). (SLIDE) FIG 5

- But analysis shows that: Area-Time Product has a minimum for some value of α (SLIDE) FIG 6

- AND THAT ENERGY / ACCESS: Also has min at some α. (SLIDE) Fig 7

- WE BELIEVE THAT THIS EMERGING THEORY CAN BE APPLIED TO "SMARTER" MEMORY STRUCTURES, TO MIXED MEM-COMP, AND MAY HELP PROVIDE A BASIS FOR A THEORY OF COSTS OF COMPUTATION IN HIGHLY CONN. SYSTEMS. MORE WHEN CARVER PROBABLY MISSES IN DEC 5