inst.eecs.berkeley.edu/~eecs251b

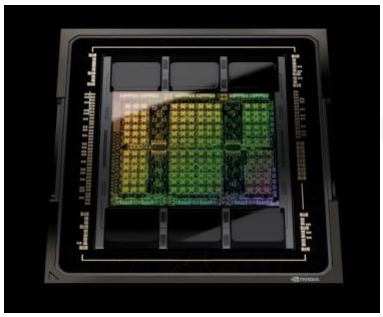
EECS251B : Advanced Digital Circuits and Systems

Lecture 18 – SRAM Peripherals

Borivoje Nikolić, Vladimir Stojanović, Sophia Shao

March 22, 2022, EETimes: Nvidia Launches Next-Gen GPU Architecture: Hopper

Nvidia unveiled its next-generation GPU architecture — named Hopper, alongside the new flagship GPU using the Hopper architecture, the H100. Perhaps surprisingly, Nvidia has not opted to go down the trendy chiplets route favored by Intel and AMD for their mammoth GPUs. While the H100 is the first GPU to use HBM3, its compute die is monolithic, 80 billion transistors in 814mm^2 built on TSMC's 4N process. Memory and compute are packaged via TSMC's CoWoS 2.5D packaging.



The new Nvidia Hopper H100 GPU – Nvidia's new flagship GPU for data center AI and scientific workloads (Source: Nvidia)

Berkeley UNIVERSITY OF CALIFORNIA

Recap

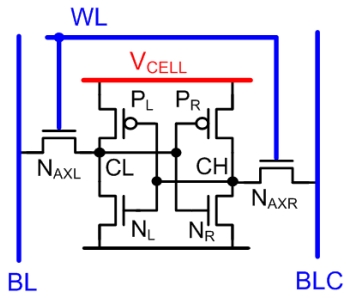
- SRAM takes a half of a die in modern chips
  - SRAM scaling is slowing down
- Static read and write margins
  - Enhanced by using assist techniques

SRAM Assist Circuits

3

Basic Ideas

- Dynamically change voltages
- Negative BL helps with writing
- Lower VDD (V_{CELL}) helps with writing
- Higher WL helps with writing, lower hurts
- Lower WL helps with read, higher hurts

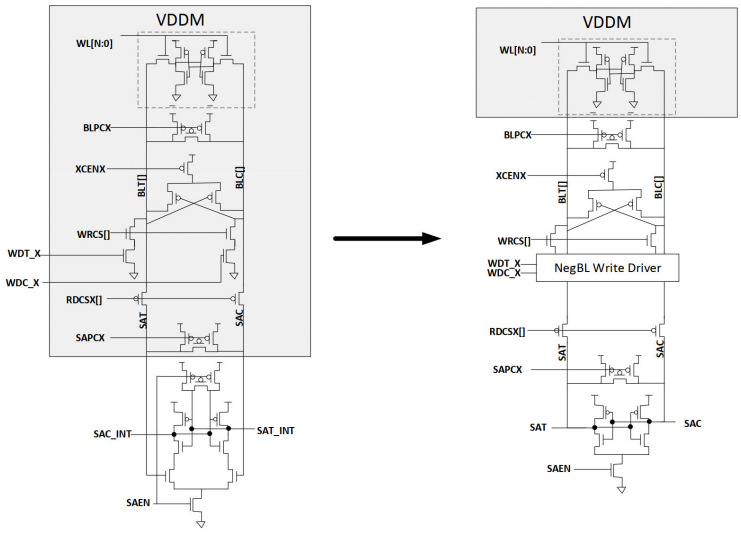


- Half-select condition: WL selected for write, but write operation is masked (BLs stay high)

4

SRAM In Practice

- 7nm AMD Zen2 (Singh, ISSCC'20)

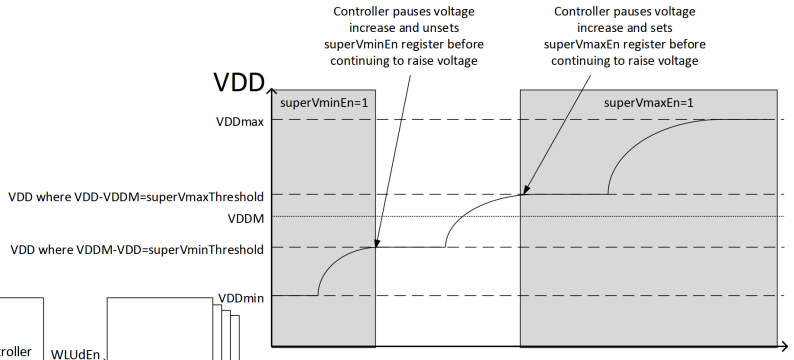


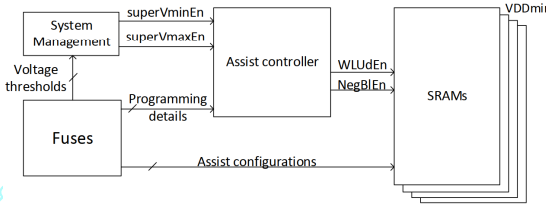
5

SRAM In Practice

- 7nm AMD Zen2 (Singh, ISSCC'20)

- Moving bitline precharge to VDD creates both bitcell stability and writeability challenges
- High level of configurability allows for silicon flexibility





6

Announcements

- Assignment 2 due on Monday
  - Quiz 2 next Tuesday

7

SRAM Peripheral Circuits

8

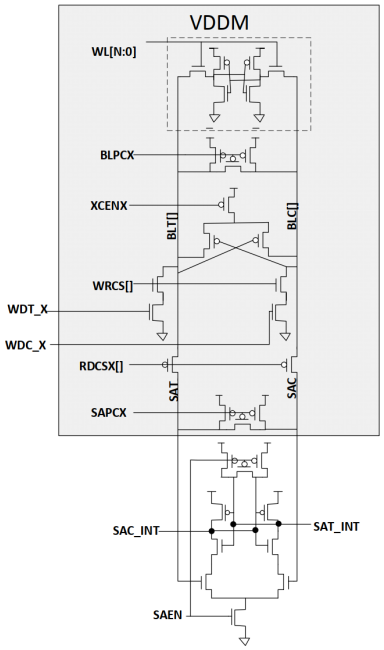
Peripheral Circuits in SRAM

- Decoders (and pre-decoders)
- Column circuitry: read, write, multiplex, mask
- Write assist techniques
- Read assist techniques
- Redundancy
- BIST
- ECC
- Power management

9

SRAM Array

- SRAM periphery:
  - Decoders (covered in EECS251A)
  - Bitline design and sense-amps

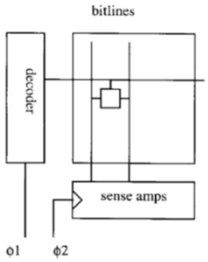


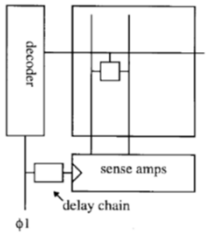
AMD Zen2
ISSCC'20

10

Sense-Amp Trigger

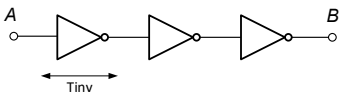
- Sense-amp trigger needs to be timed carefully
  - Too early: Incorrect evaluation
  - Too late: Unnecessary timing margin



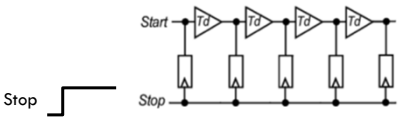


Aside: Delay Lines, Replicas and Time Amplification

- We will encounter it several times in this course
  - Used in a wide range of mixed-signal circuits
- A simple delay line



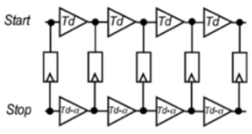
Time-to-digital converter (TDC)



Start-Stop difference read out as a thermometer-coded binary value

Resolution set by inverter delay

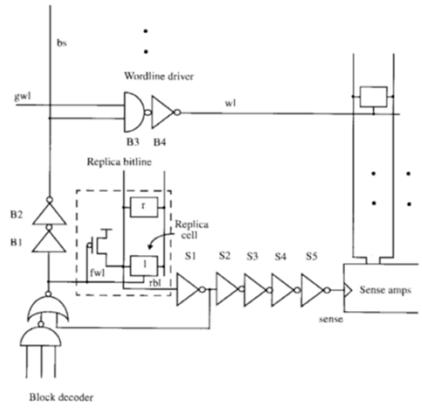
Sub-inverter delays are hard to generate
Small α requires large area

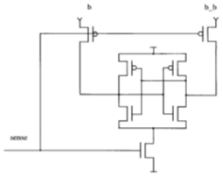


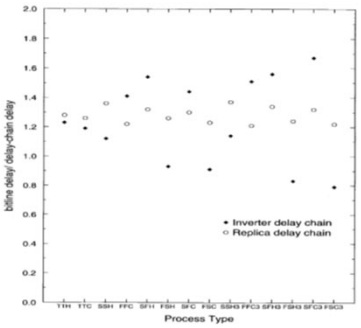
Lee, Abidi, JSSC 4/08

Sense-Amp Triggering

- Replica bitline







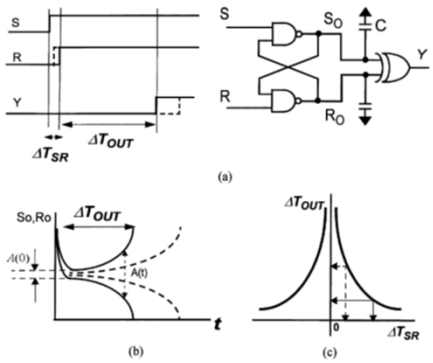
Replica delay tracks better across corners
But still mistracks across a wide range of supplies

Amrutur, Horowitz, JSSC 8/98

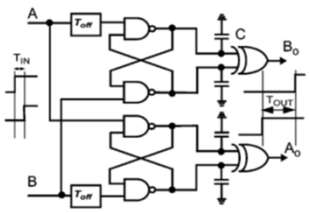
13

Time Amplification

- Time amplified through metastability (by using setup time characteristics)



Time amplifier



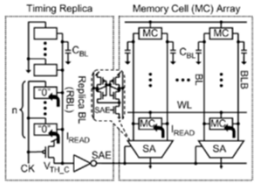
$T_{out} > T_{in}$,
Adjustable by T_{off}, C

Lee, Abidi, JSSC 4/08

14

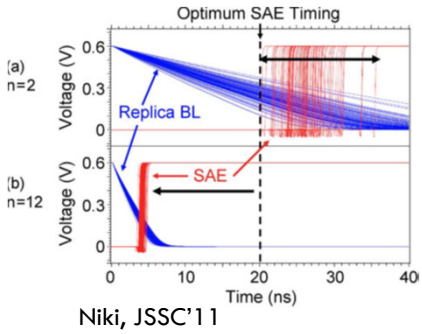
Voltage Scaling: Multiplicative Replica Bitline

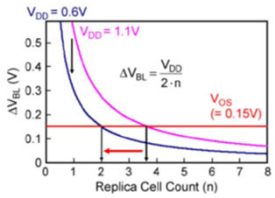
- Conventional replica



n replica cells discharging replica BL in parallel to reduce the current variation by $\sqrt{n}$

Threshold for discharge is set accordingly to $V_{DD} - nV_{os}$
Limits n to ~2-4



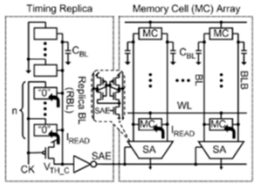


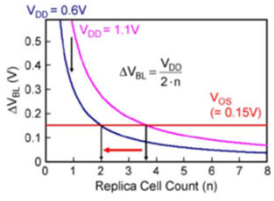
Niki, JSSC'11

15

Voltage Scaling: Multiplicative Replica Bitline

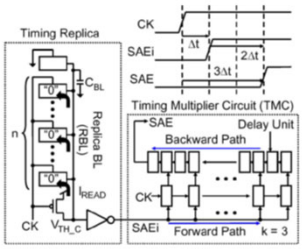
- Conventional replica

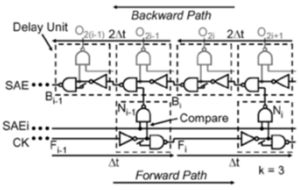




- Programmable replica delay
- Multiplicative replica scales the delay, w/o increasing variance

- Multiplictive replica





Niki, JSSC'11

16

Redundancy and ECC

17

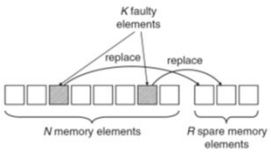
Redundancy and ECC

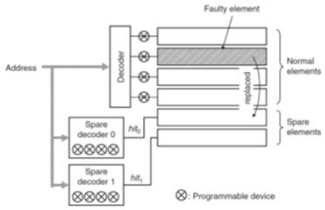
- Redundancy
  - Spare columns (or rows)
  - Selected at test via eFuse
  - Possible to dynamically program redundancy
- ECC
  - Error detection/correction codes
  - Parity
  - SECDED
  - DECTED

18

Redundancy

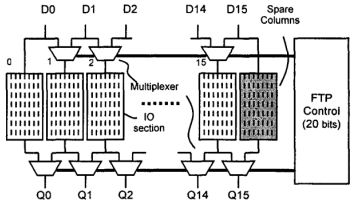
- Principle
  - Rows





Horiguchi, Itoh, Springer 2011.

  - Columns



McPartland, CICC'00.

Redundancy

- Effectiveness (Bickford, 2008)

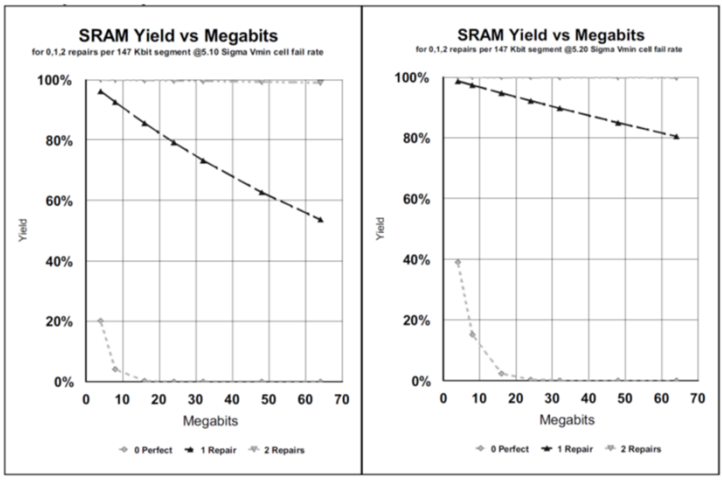


Figure 1: Modeled Yield impact comparison for 65 nm SRAM complier. Vmin cell fail rate used in analysis shown in the left chart is 5.10 sigma. Vmin cell fail rate used in the analysis shown in the right chart is 5.20 sigma. 147 Kbit segment is a standardized array size block segment used for comparison purposes

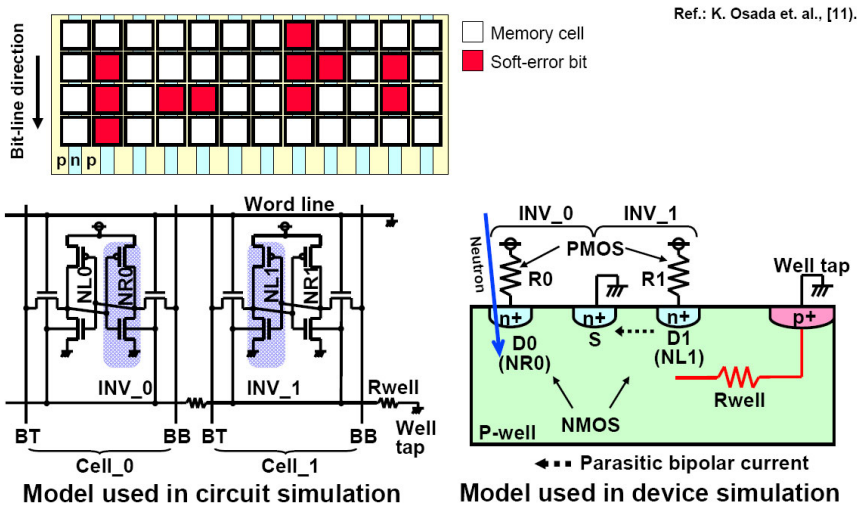
Soft Errors

- From packaging and cosmic rays
- Packaging:
  - Lead ore contains Po-210 -> (5 days) -> Bi-210 -> (22.3 years) -> Pb-210
  - Or Po-210 -> (138.4 days) -> Pb-210
  - Need 'old lead'
- Cosmic rays
  - Large particles collide with Earth's atmosphere to produce alpha (and other) particles

Error Correction

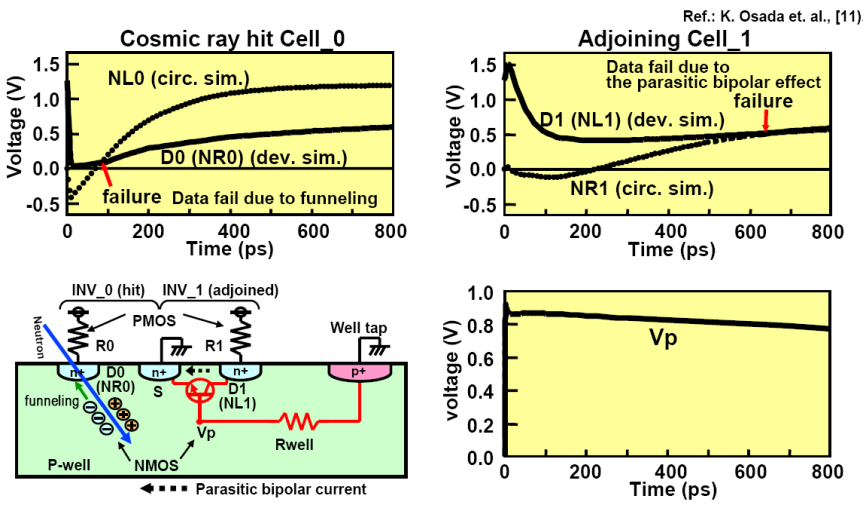
- Parity (SED)
  - $p = d_7 \oplus d_6 \oplus d_5 \oplus d_4 \oplus d_3 \oplus d_2 \oplus d_1 \oplus d_0$
- SECDED
- DECTED

Multi-bit Errors



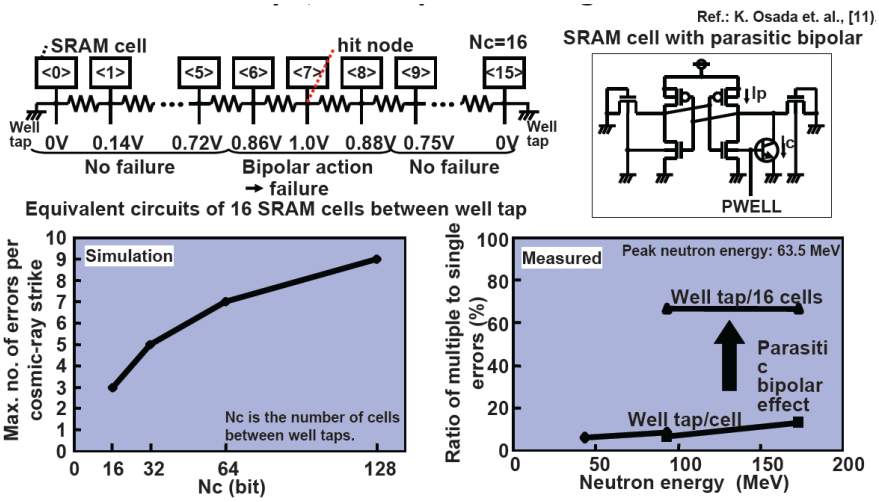
Kawahara, ISSCC'07 tutorial

Multi-bit Errors



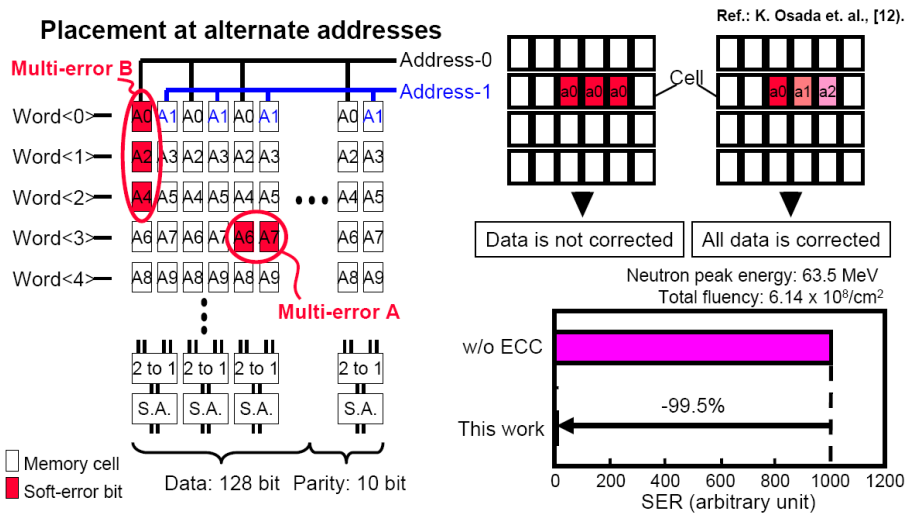
Kawahara, ISSCC'07 tutorial

Multi-bit Errors



25

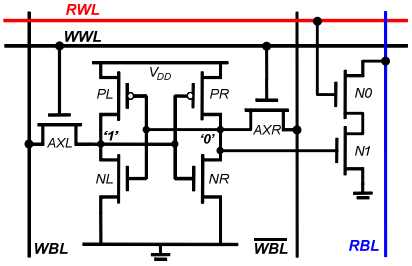
Multi-bit Errors: Interleaving



26

6T SRAM Alternatives

8-SRAM



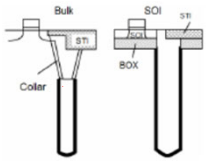
- Read circuit?
- Interleaving?

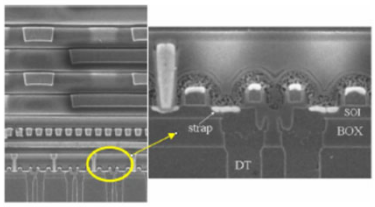
- Dual-port read/write capability (register-file-like cells)
- N0, N1 separates read and write
  - No Read SNM constraint
  - Half-selected cells still undergo read
- Stacked transistors reduce leakage

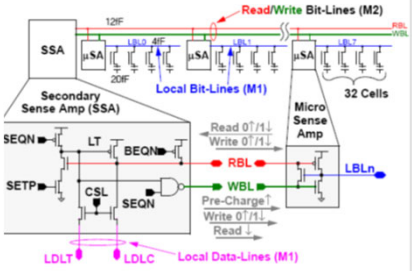
L. Chang, *VLSI Circuits* 2005

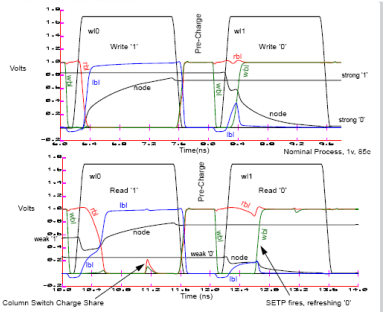
eDRAM

- Process cost: Added trench capacitor









Barth, ISSCC'07, Wang, IEDM'06

Crosspoint Memories

- Barrett, IRE Trans. Comp. 1961.

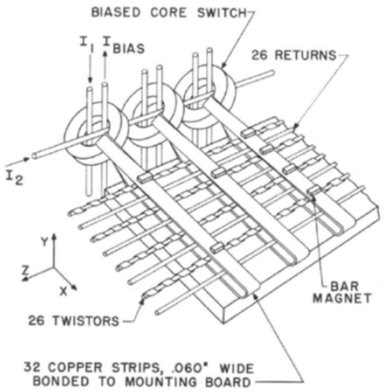
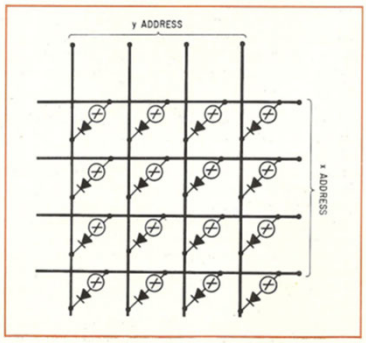


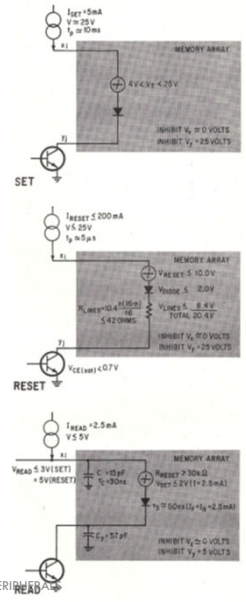
Fig. 2—Memory structure. I_1 and I_2 are access drive currents to core-selection switch. Presence or absence of a magnet over a twistor-strip solenoid crosspoint yields a "zero" or "one." Signals observed between twistor and return wire.

Crosspoint Memories

- Neale, Nelson, Moore, Electronics'70
  - 16 x 16 array (256b) of 'read-mostly memory'

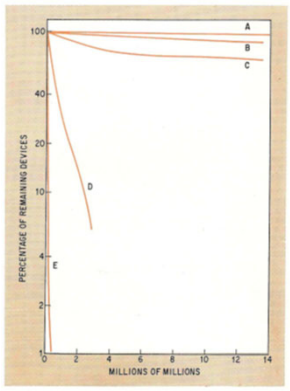


Crosspoint Memory



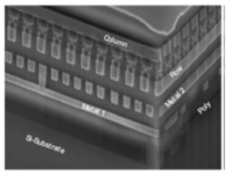
- Four modes
  - Form
  - Set
  - Reset
  - Read

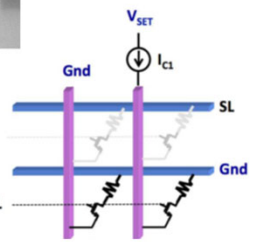
› **Endurance**

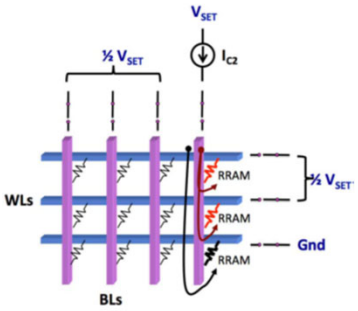


3D Crosspoint Arrays

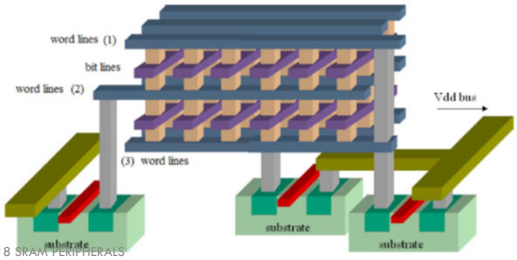
- Kau, IEDM'09







- Yeh, JSSC'15

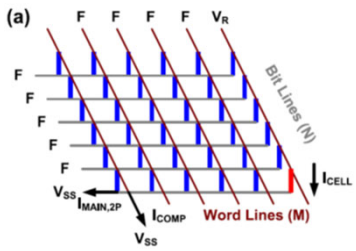


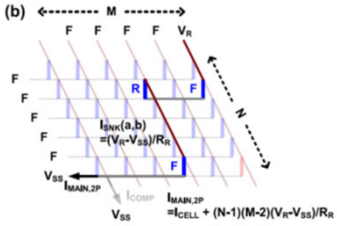
- Ou, JSSC'11

33

Crosspoint Arrays

- Read and sneak currents





Bae, TED 4/17

34

Summary

- SRAM periphery
  - Decoders
  - Assist circuits
  - Sense amp timing replicas
- 6-T SRAM alternatives
  - 8-T SRAM
  - eDRAM
  - Crosspoint arrays (e.g. RRAM)

Next Lecture

- Low-power design
  - Power-performance tradeoffs