inst.eecs.berkeley.edu/~eecs251b

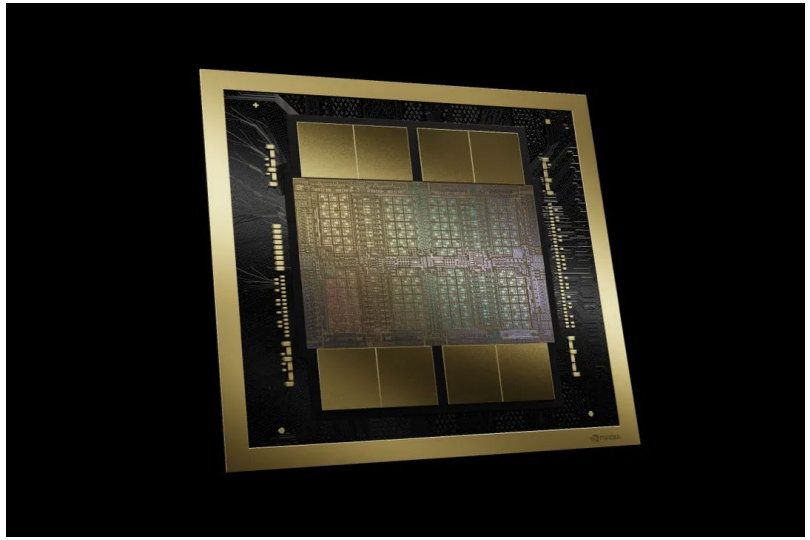
EECS251B : Advanced Digital Circuits and Systems

Lecture 19 – SRAM

Borivoje Nikolić

NVIDIA Announces New GPU Architecture

March 17, 2024. Blackwell-architecture GPUs pack 208 billion transistors and are manufactured using a custom-built TSMC 4NP process. All Blackwell products feature two reticle-limited dies connected by a 10 terabytes per second (TB/s) chip-to-chip interconnect in a unified single GPU.



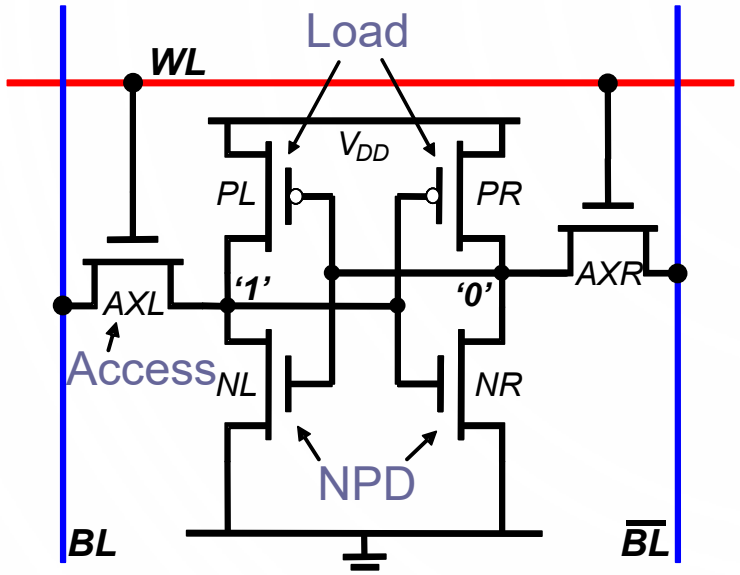
https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/

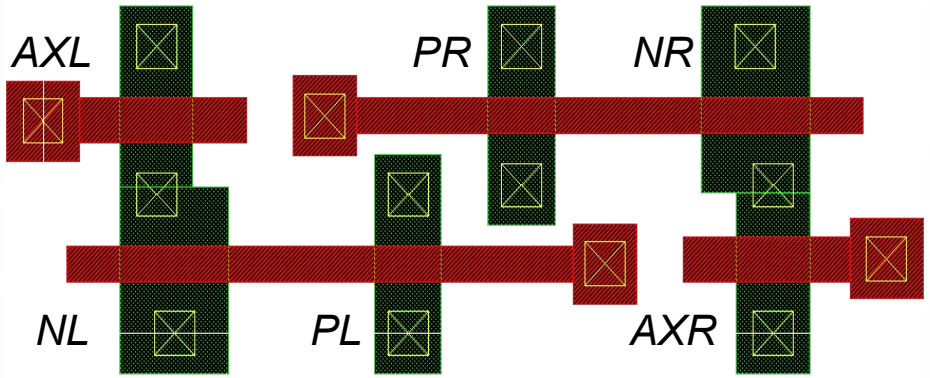
Announcements

- Project
  - Midterm reports due tomorrow!
  - Preliminary design review after Spring break
- Homework 3 due tomorrow
  - Quiz 3 after Spring break
- Lab 5 posted today

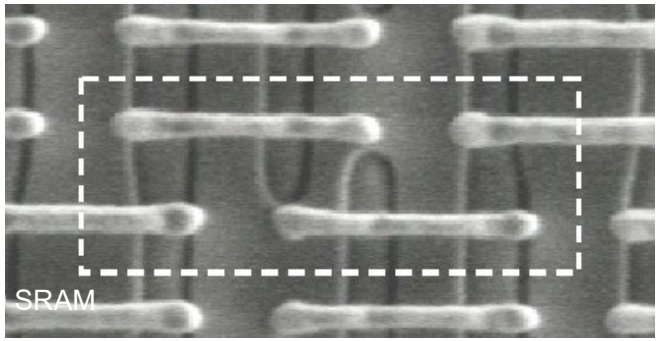
SRAM

6-T SRAM Cell



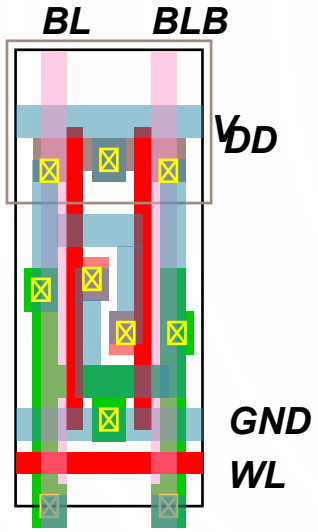


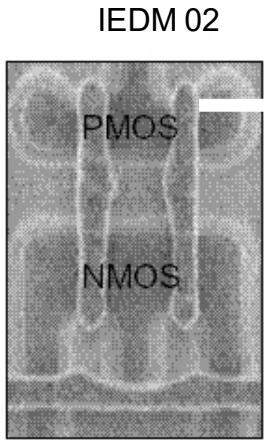
Long Cell Topology



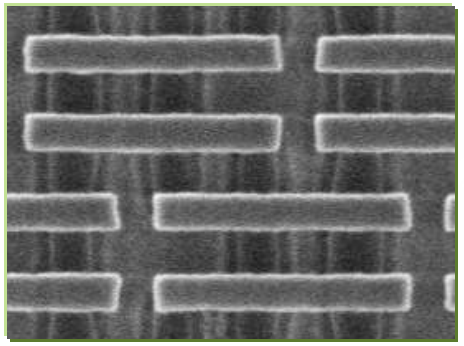
$$(W/L)_{NL} > (W/L)_{AXL}$$

SRAM Cell Design Trends

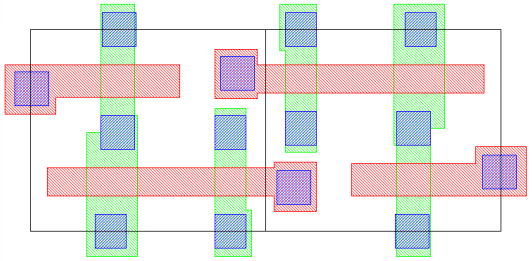




Cell in 90nm
(1μm²)



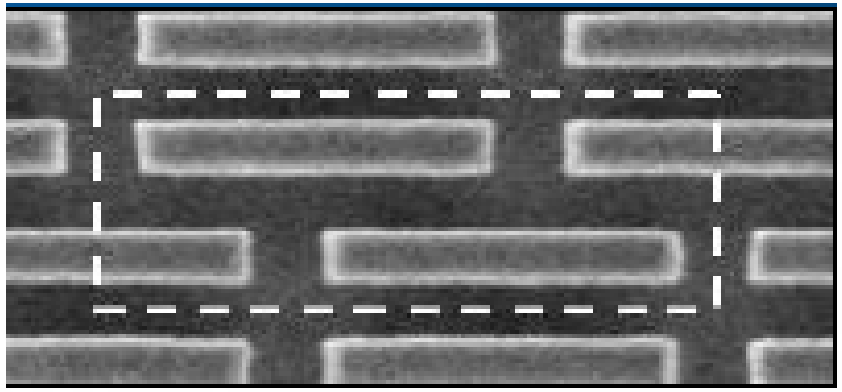
Cell in 32nm
(0.171μm²)



- **Key enabling technology: STI**
- **Impact: Increased cell density**

SRAM Cell Trends (22nm)

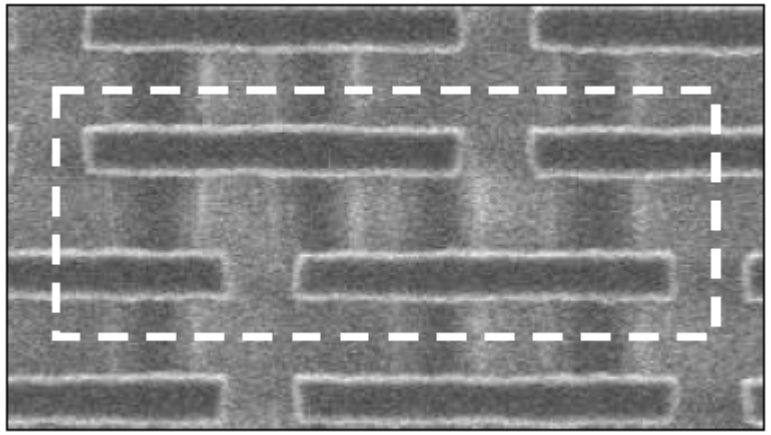
finFET – integer number of fins



$0.092\mu m^2$ cell in 22nm from Intel (IDF'09)

A little analysis by using a ruler:
- Aspect ratio 2.9
- Height ~178nm, Width ~518nm
- Gate ~ 45nm (Lg is smaller for logic)

planar

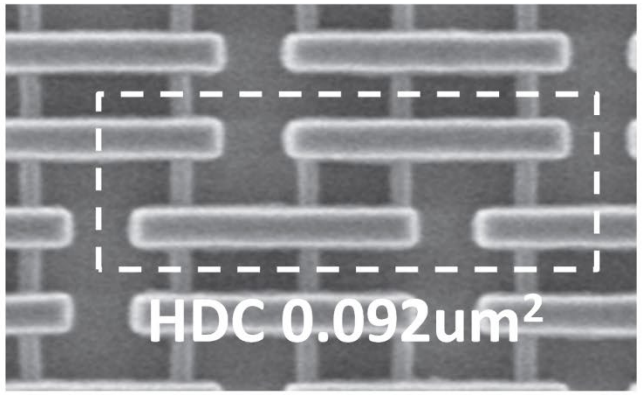


$0.346\mu m^2$ cell in 45nm from Intel (IEDM'07)

22nm SRAM – Discrete Widths

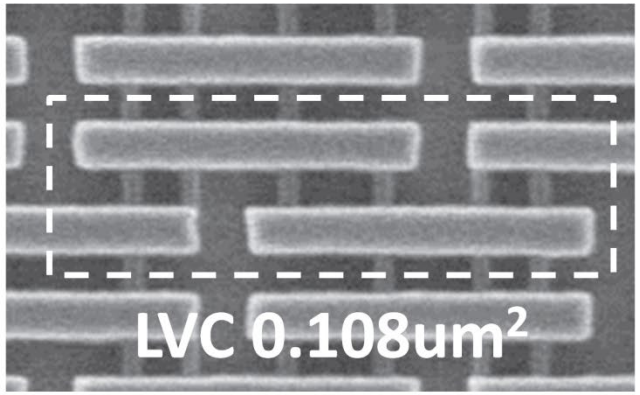
- FinFET cell design

High-Density Cell

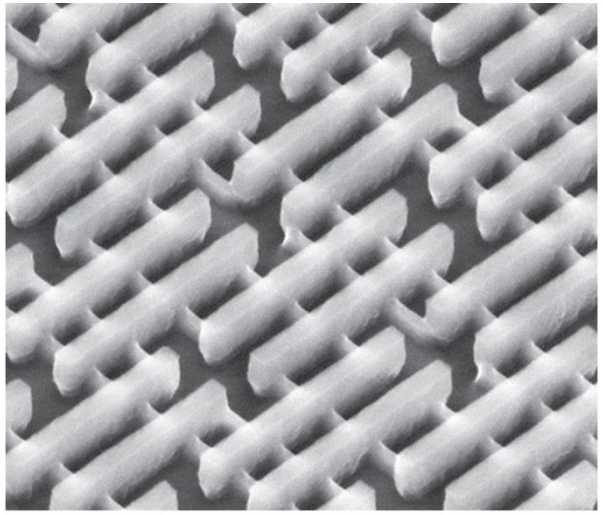


(PD:PG:PU)

Low-Voltage Cell

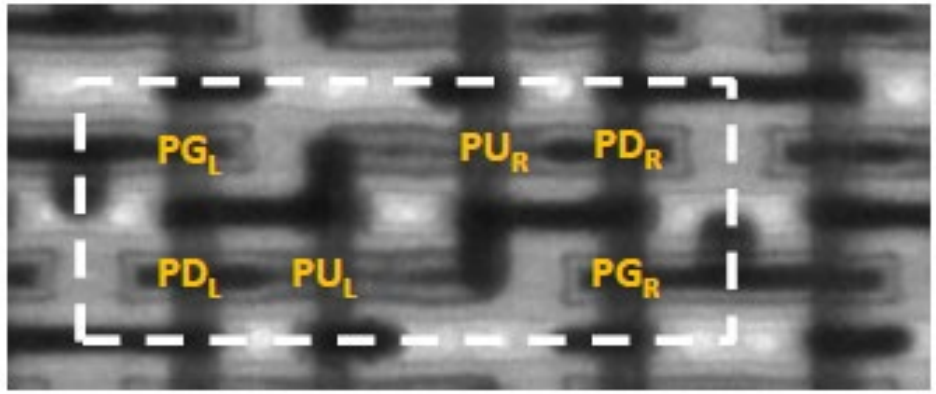


(PD:PG:PU)



E. Karl, ISSCC'12

14nm SRAM

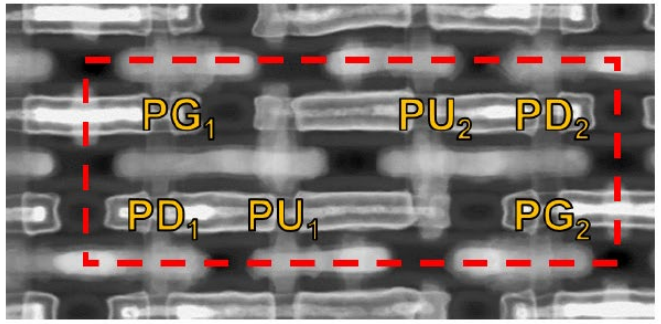


- Aspect ratio ~2.5
- Cell area = 0.05um^2
  - Height = 140nm (2 gate p)
  - Width = 350nm
  - Lg ~ 32nm (longer than for logic)

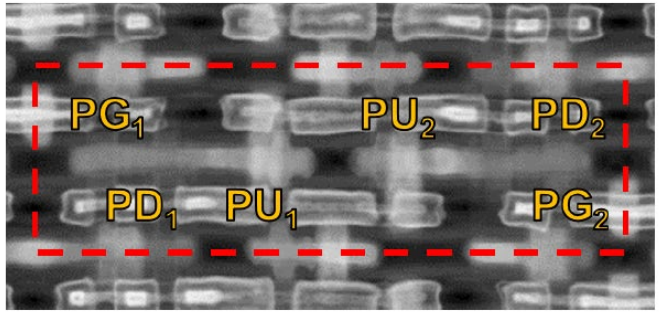
E. Karl, ISSCC'15

10nm SRAM

HDC
0.0312 μm²



LVC
0.0367 μm²

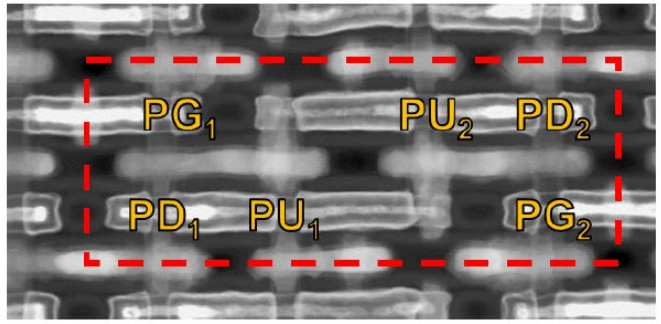


- High-Density Cell (HDC)
  1:1:1 (PU:PG:PD)
- Low-Voltage Cell (LVC)
  1:1:2 (PU:PG:PD)

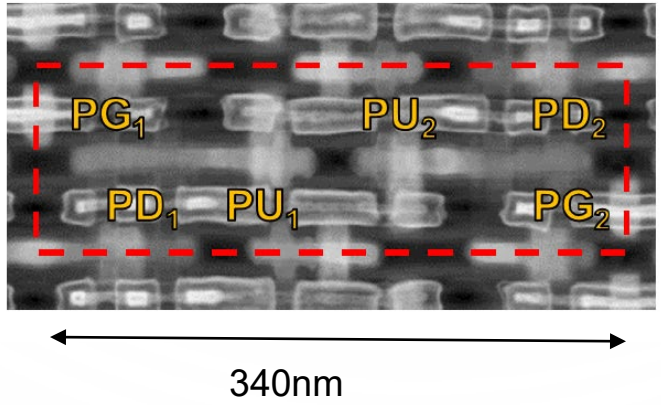
Guo, ISSCC'18

10nm SRAM + Ruler

HDC
0.0312 μm²



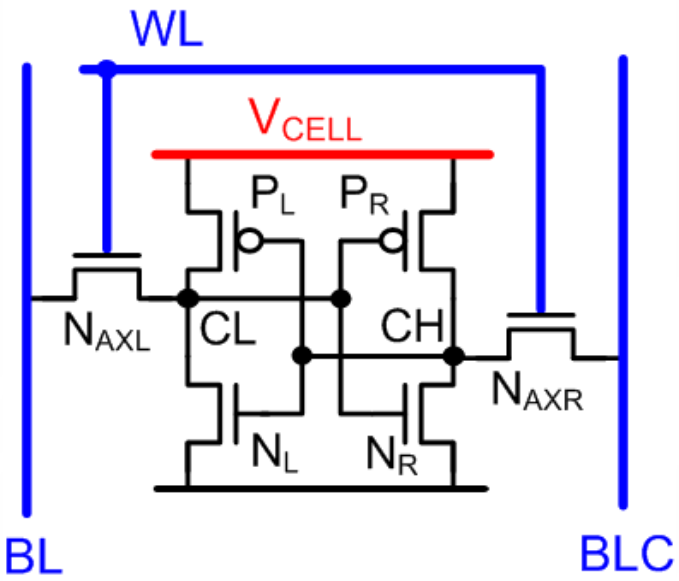
LVC
0.0367 μm²



SRAM: Assist Circuits

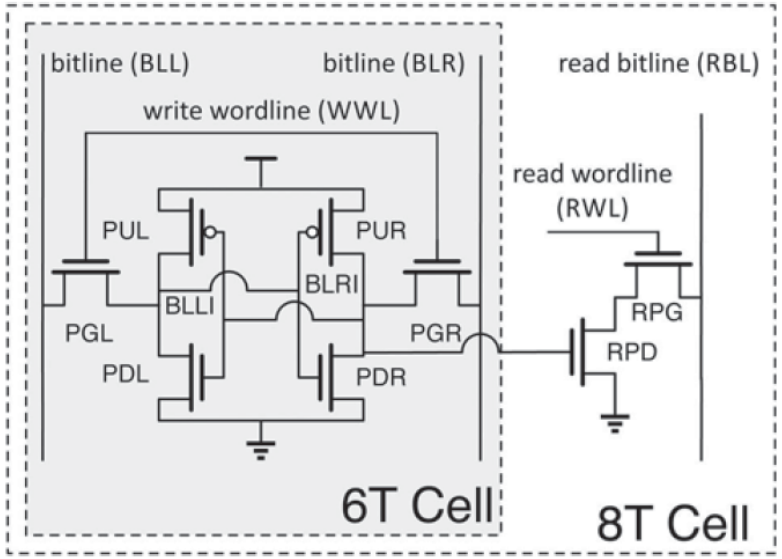
Basic Ideas

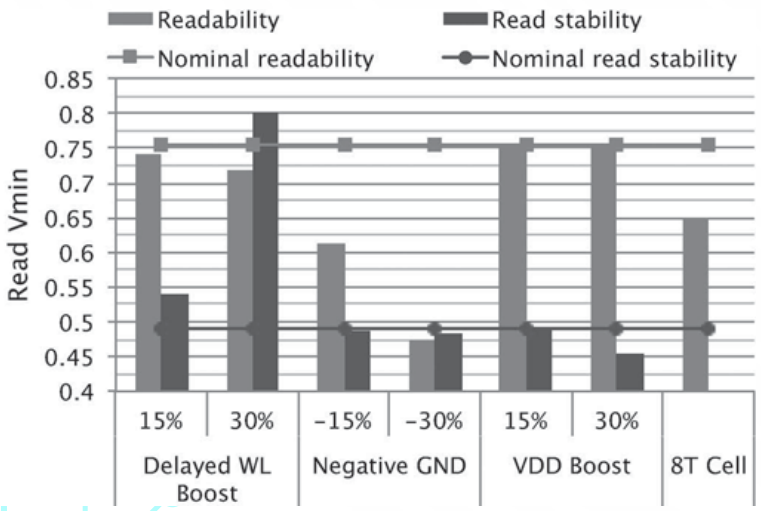
- Dynamically change voltages
- Negative BL helps with writing
- Lower VDD (V_{CELL}) helps with writing
- Higher WL helps with writing, lower hurts
- Lower WL helps with read, higher hurts

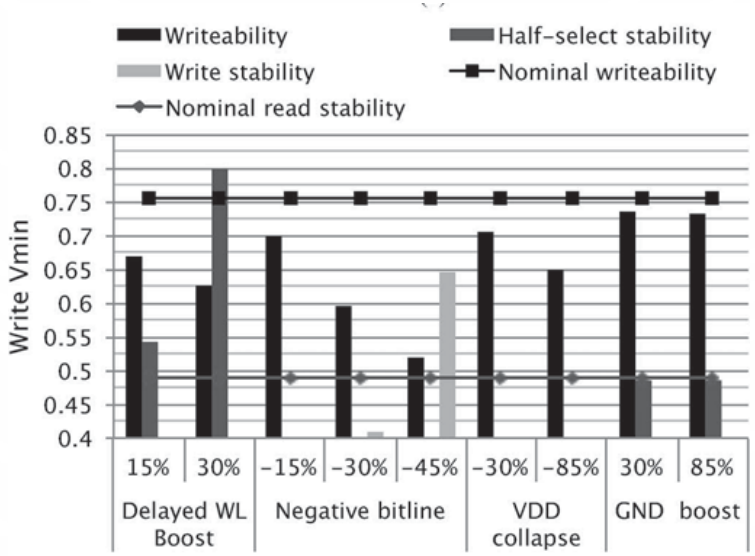


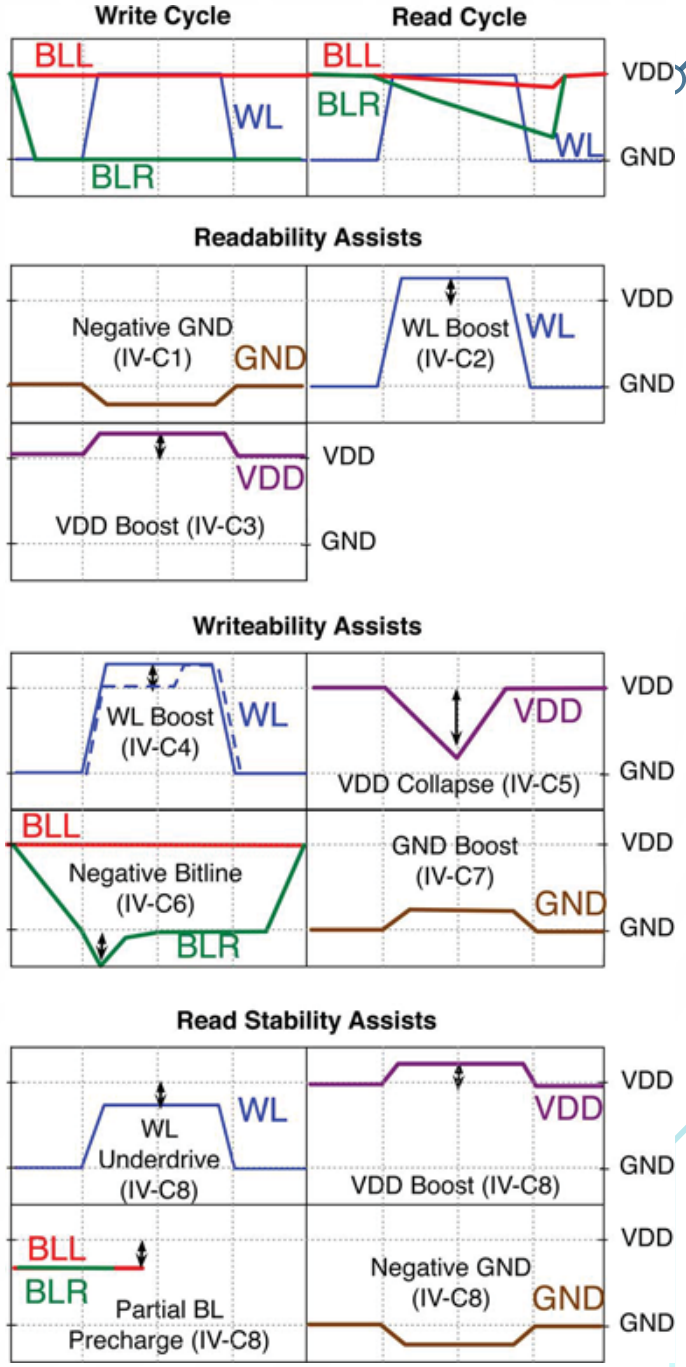
- Half-select condition: WL selected for write, but write operation is masked (BLs stay high)

Impact on performance





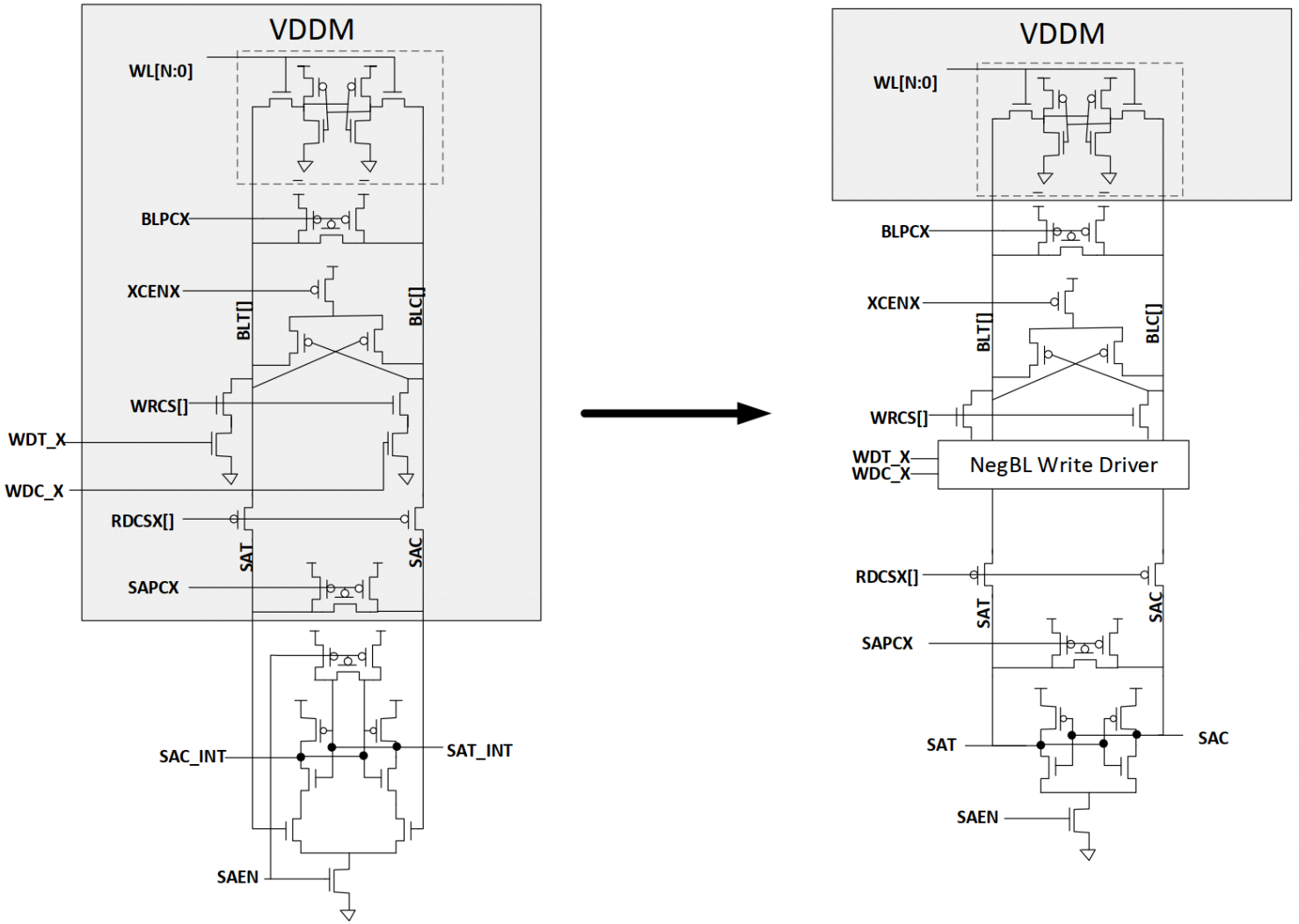




Zimmer, TCAS 2012

SRAM In Practice

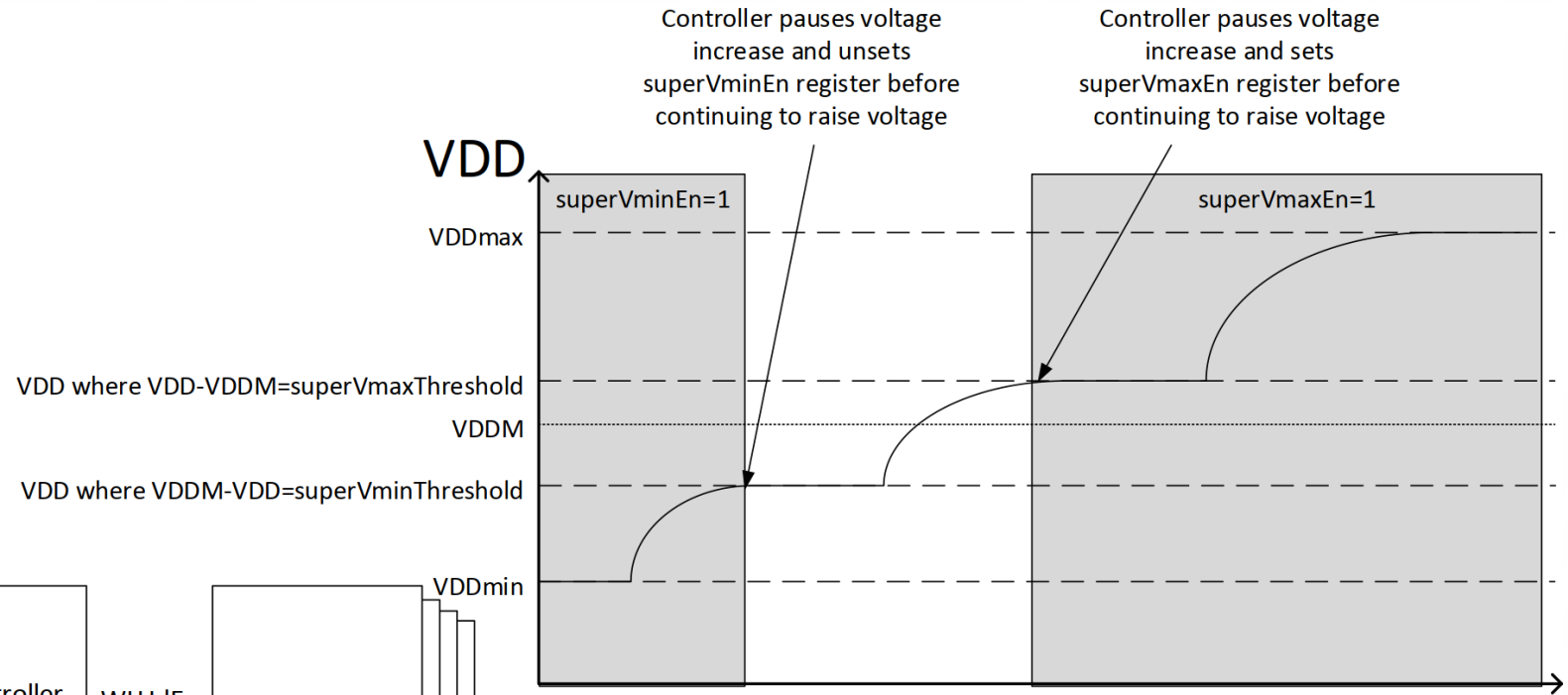
- 7nm AMD Zen2 (Singh, ISSCC'20)

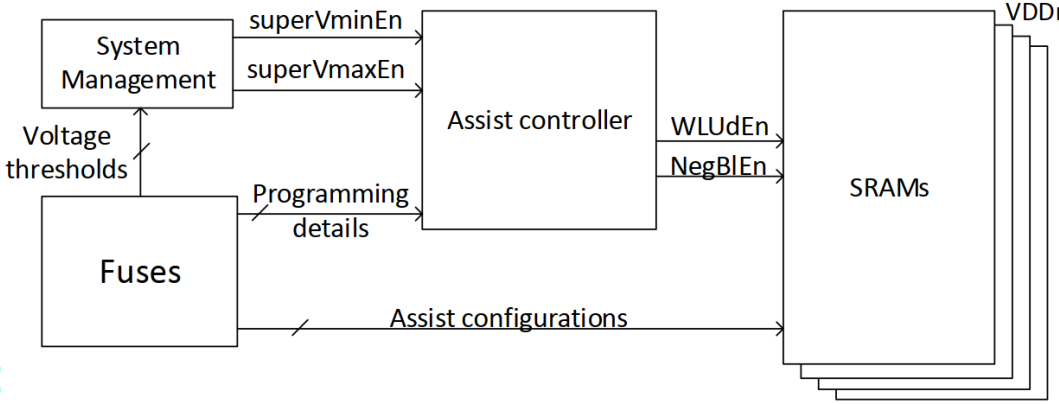


SRAM In Practice

- 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





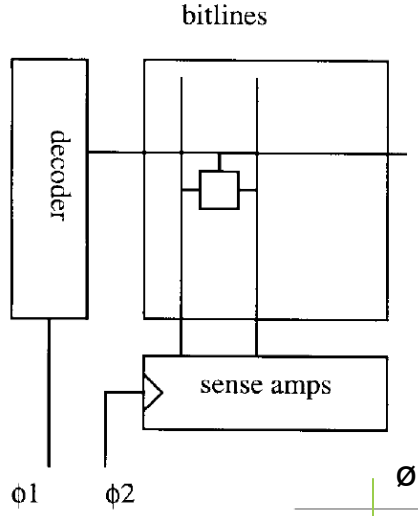
SRAM Peripheral Circuits

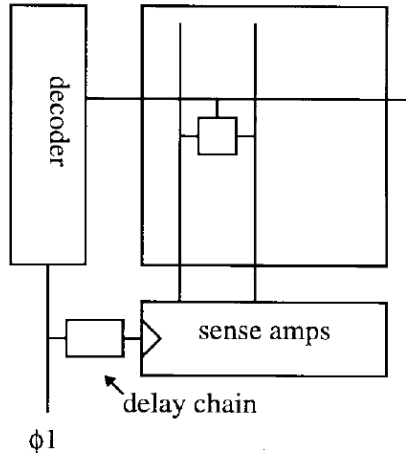
Peripheral Circuits in SRAM

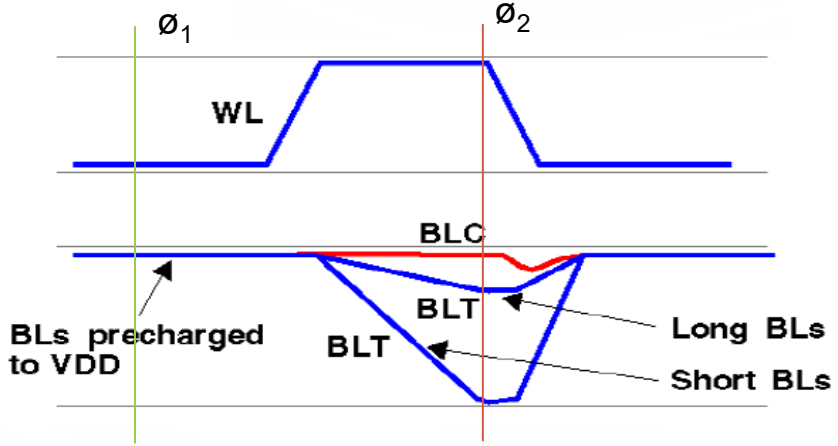
- Decoders (and pre-decoders)
- Column circuitry: read, write, multiplex, mask
- Write assist techniques
- Read assist techniques
- Redundancy
- BIST
- ECC
- Power management

Sense-Amp Trigger

- Sense-amp trigger needs to be timed carefully
  - Too early: Incorrect evaluation
  - Too late: Unnecessary timing margin
- Problem: Delay based on inverter chains does not track the delay of the memory cell

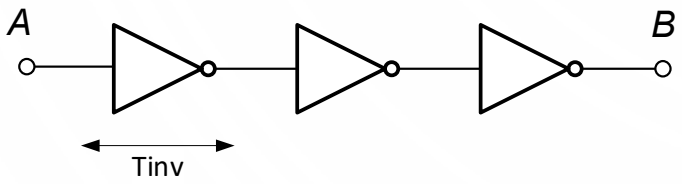




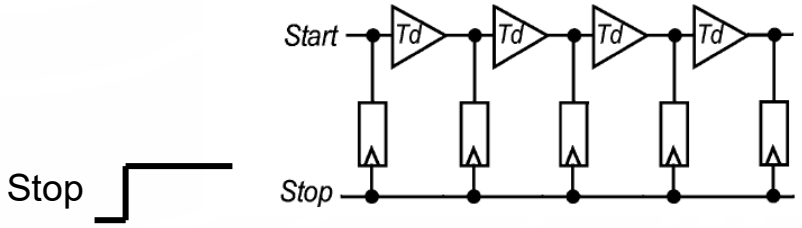


Aside: Delay Lines, Replicas and Time Amplification

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



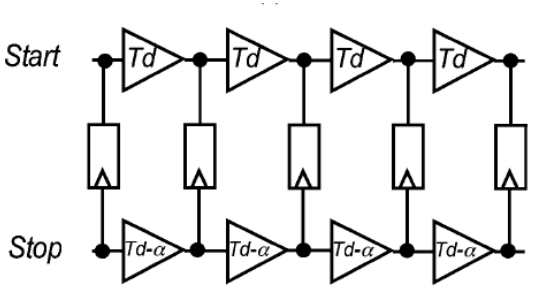
Time-to-digital converter (TDC)



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

Resolution set by inverter delay

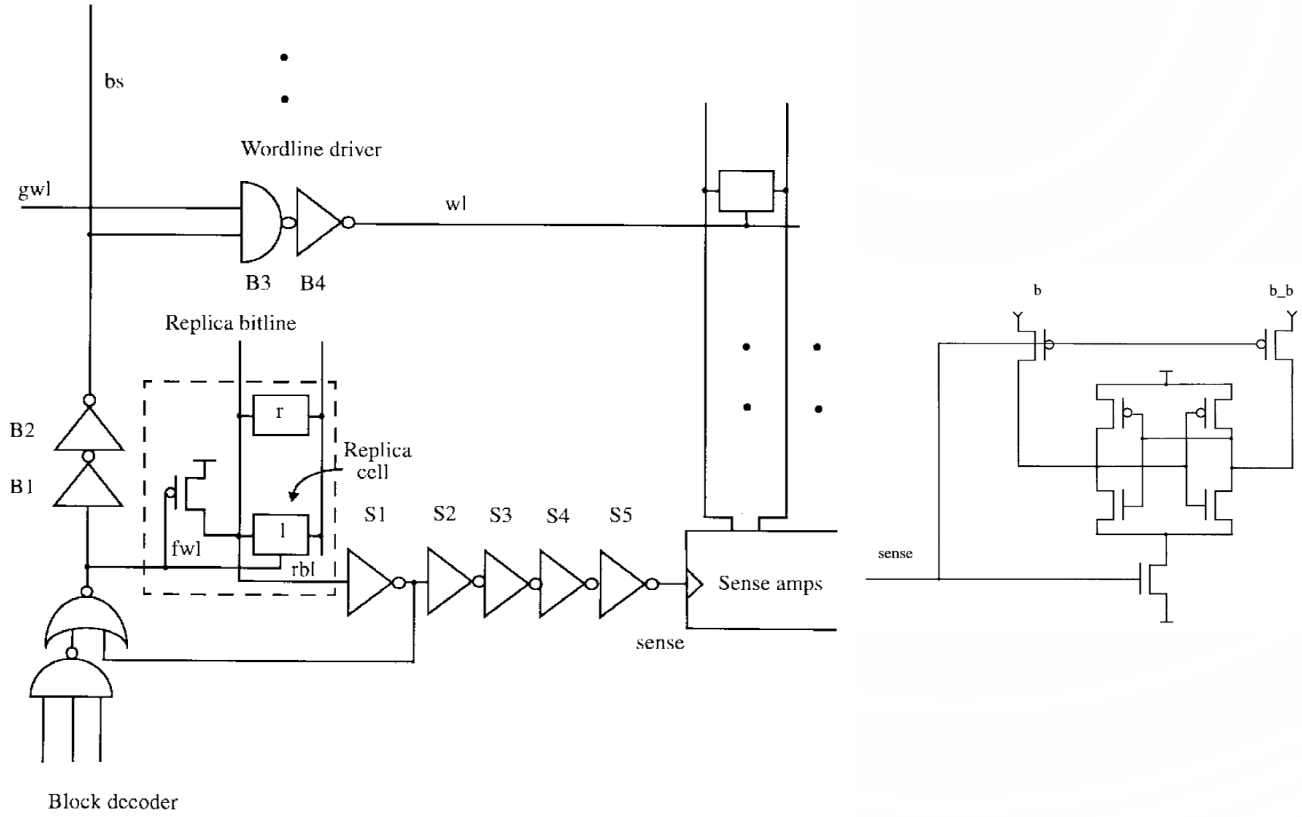
Sub-inverter delays are hard to generate
Small α requires large area

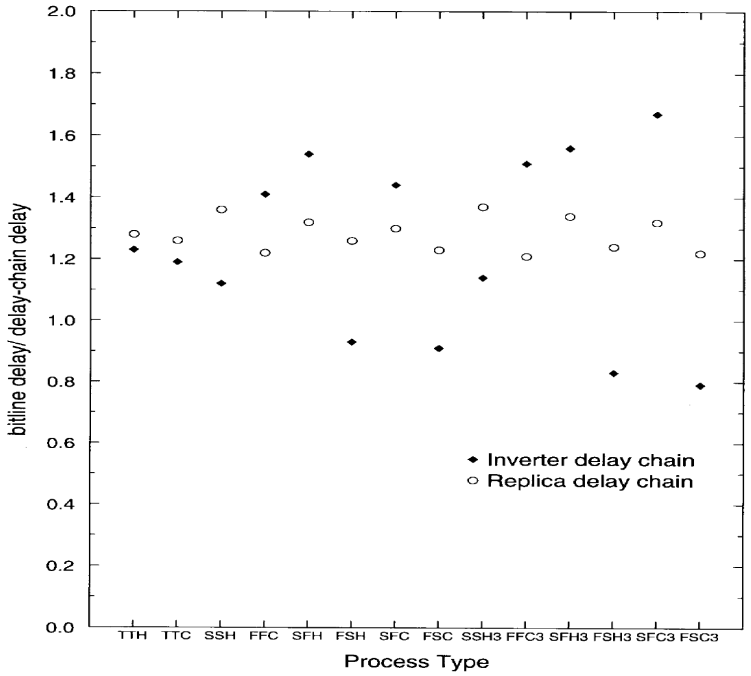


Lee, Abidi, JSSC 4/08

Sense-Amp Triggering

- Replica bitline





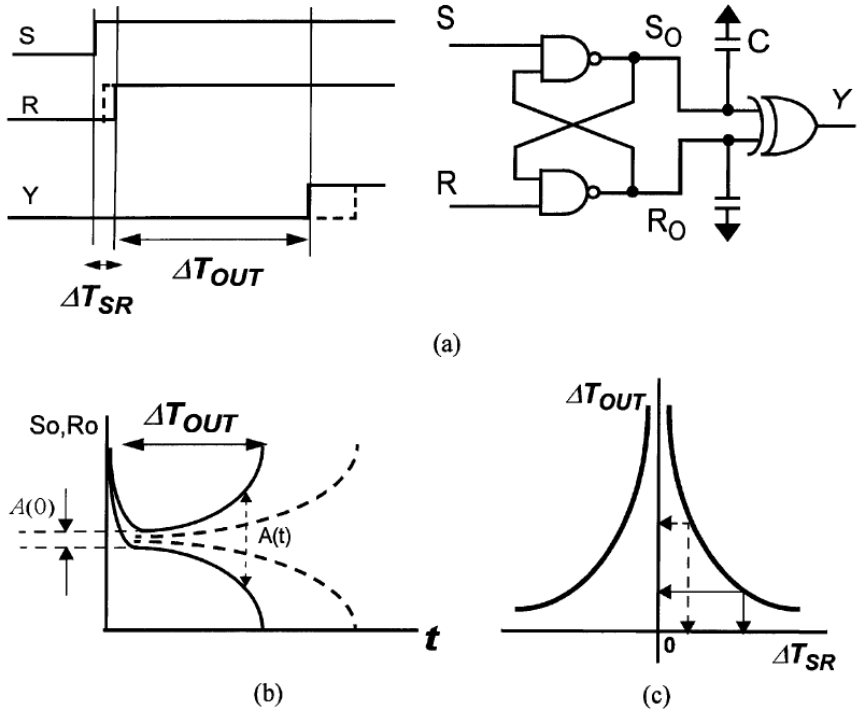
Replica delay tracks better across corners

But still mistracks across a wide range of supplies

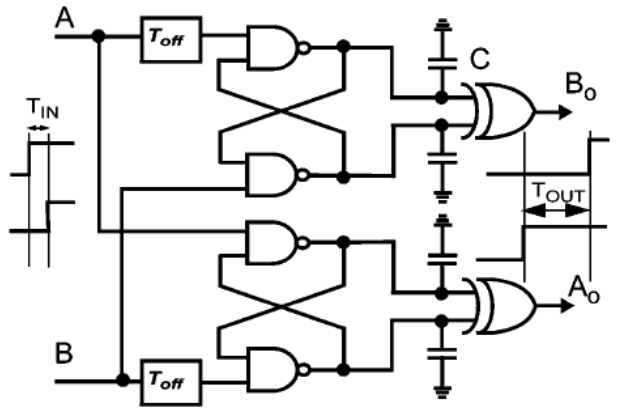
Amrutur, Horowitz, JSSC 8/98

Time Amplification

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



Time amplifier

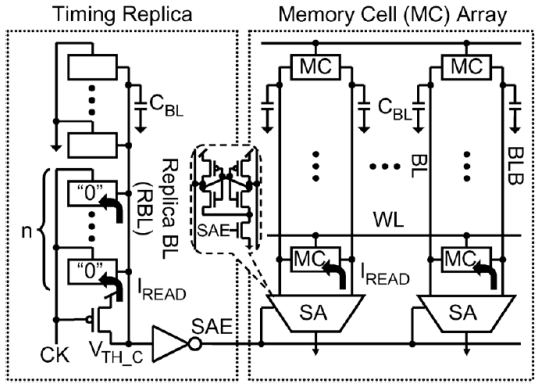


$T_{out} > T_{in}$,
Adjustable by T_{off}, C

Lee, Abidi, JSSC 4/08

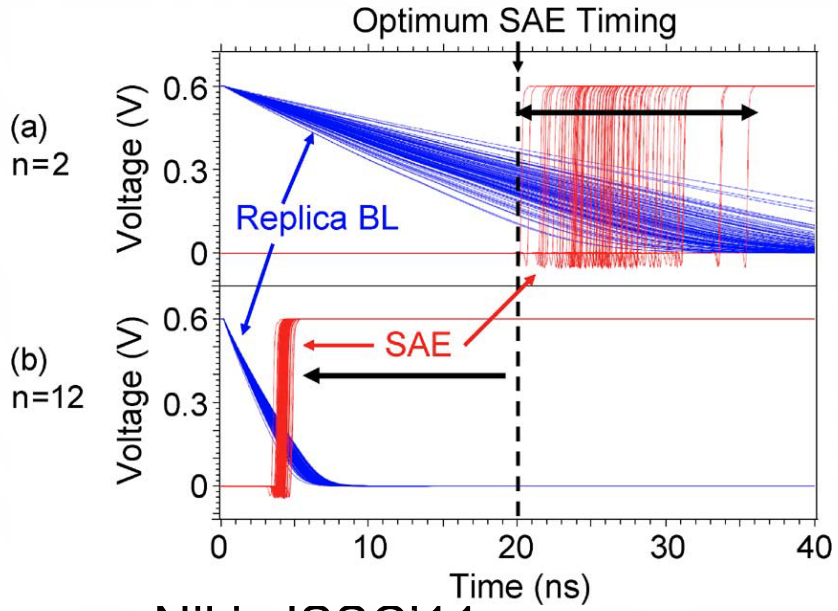
Voltage Scaling: Multiplicative Replica Bitline

- Conventional replica

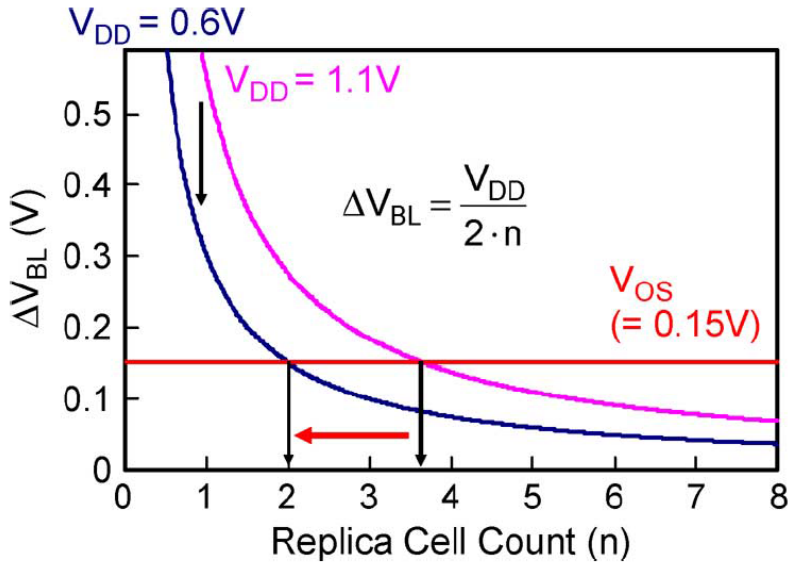


n replica cells discharging replica BL in parallel to reduce the current/cell variation by $\sqrt{n}$

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

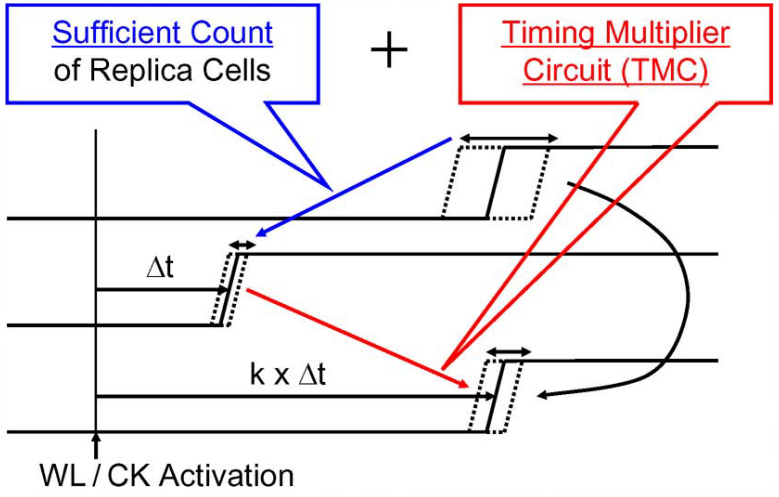


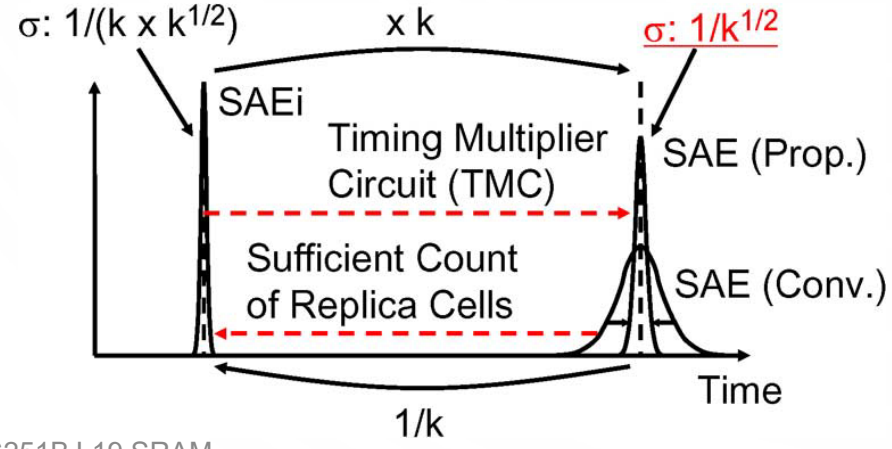
Niki, JSSC'11



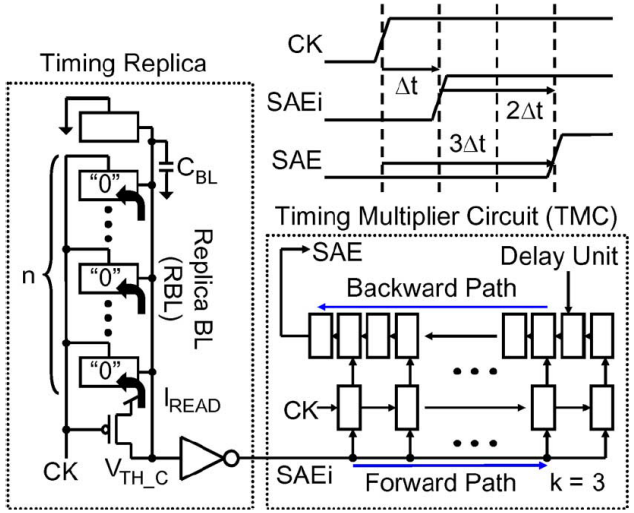
Voltage Scaling: Multiplicative Replica Bitline

- Multiplicative replica

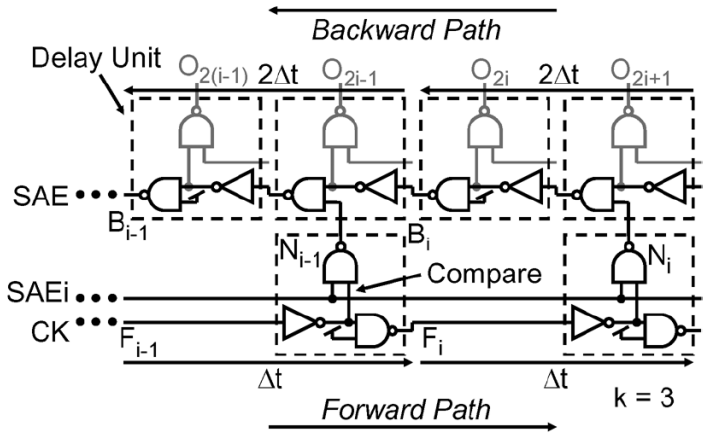




- Programmable replica delay
- Multiplicative replica scales the delay, w/o increasing variance correspondingly



Forward path digitizes SAEi to CK delay
Backward path multiplies



Niki, JSSC'11

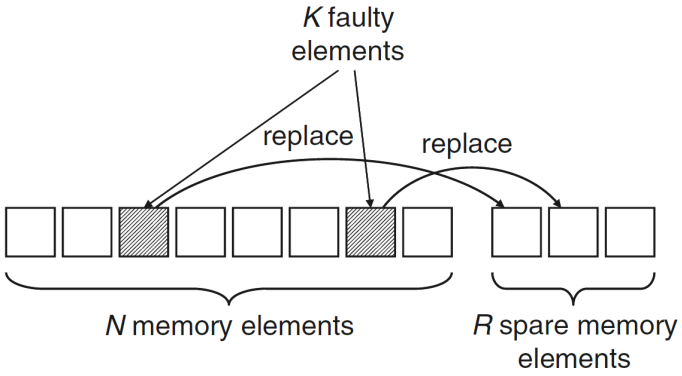
Redundancy and ECC

Redundancy and ECC

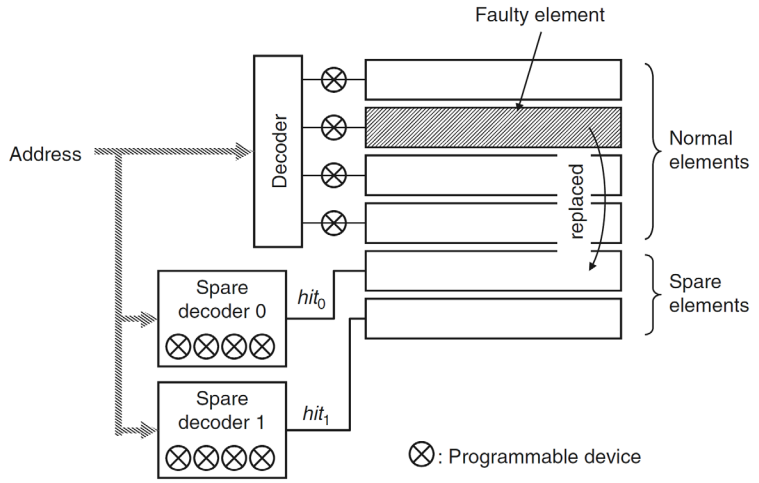
- Redundancy
  - Spare columns (or rows)
  - Selected at test via eFuse
  - Possible to dynamically program redundancy
- ECC
  - Error detection/correction codes
  - Parity
  - SECDED
  - DECTED

Redundancy

- Principle

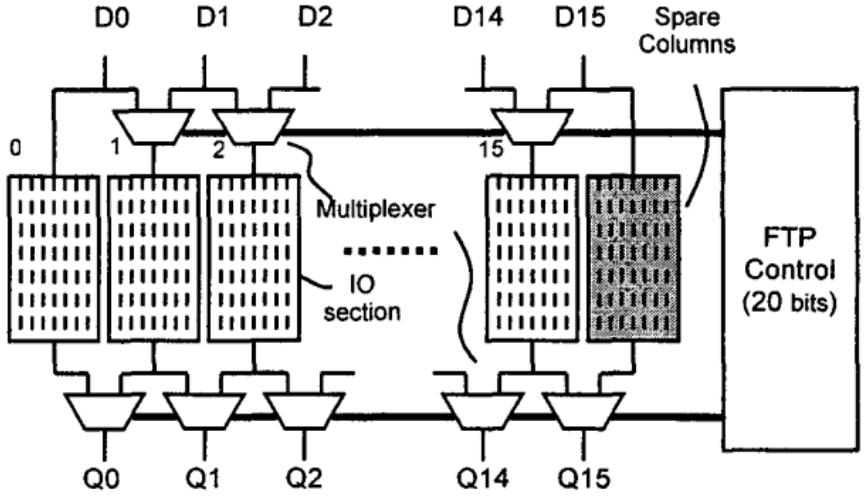


- Rows



Horiguchi, Itoh, Springer 2011.

- Columns



McPartland, CICC'00.

Redundancy

- Effectiveness (Bickford, 2008)

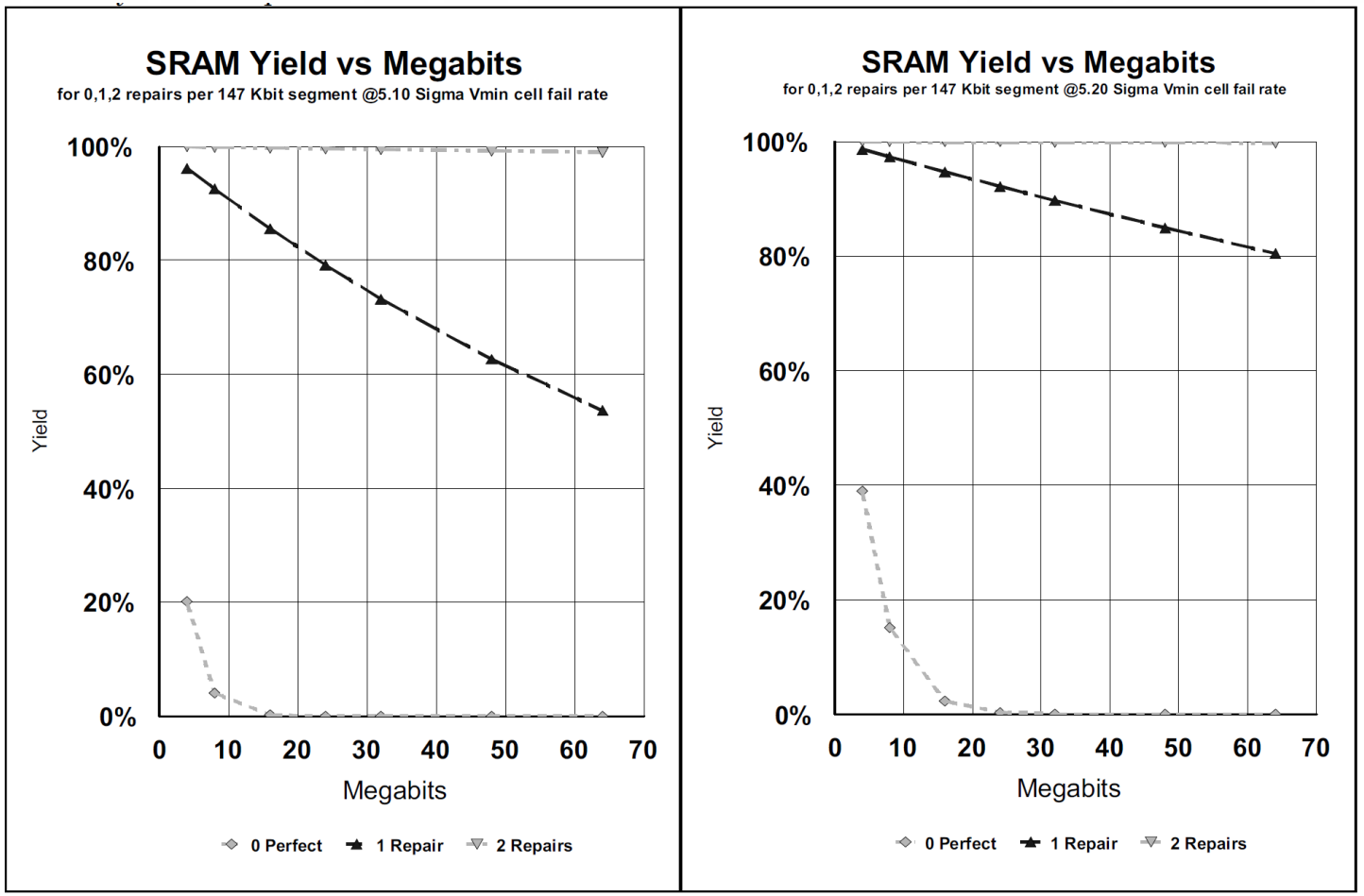


Figure 1: Modeled Yield impact comparison for 65 nm SRAM complier. Vmin cell fail rate used in analysis shown in the left chart is 5.10 sigma. Vmin cell fail rate used in the analysis shown in the right chart is 5.20 sigma. 147 Kbit segment is a standardized array size block segment used for comparison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

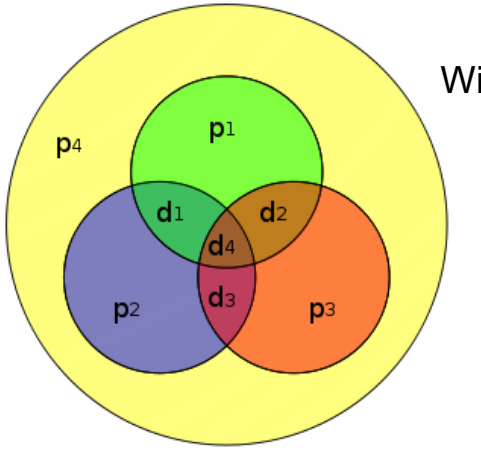
Soft Errors

- 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 – Single Error Detection (SED)
  - $p = d_7 \oplus d_6 \oplus d_5 \oplus d_4 \oplus d_3 \oplus d_2 \oplus d_1 \oplus d_0$

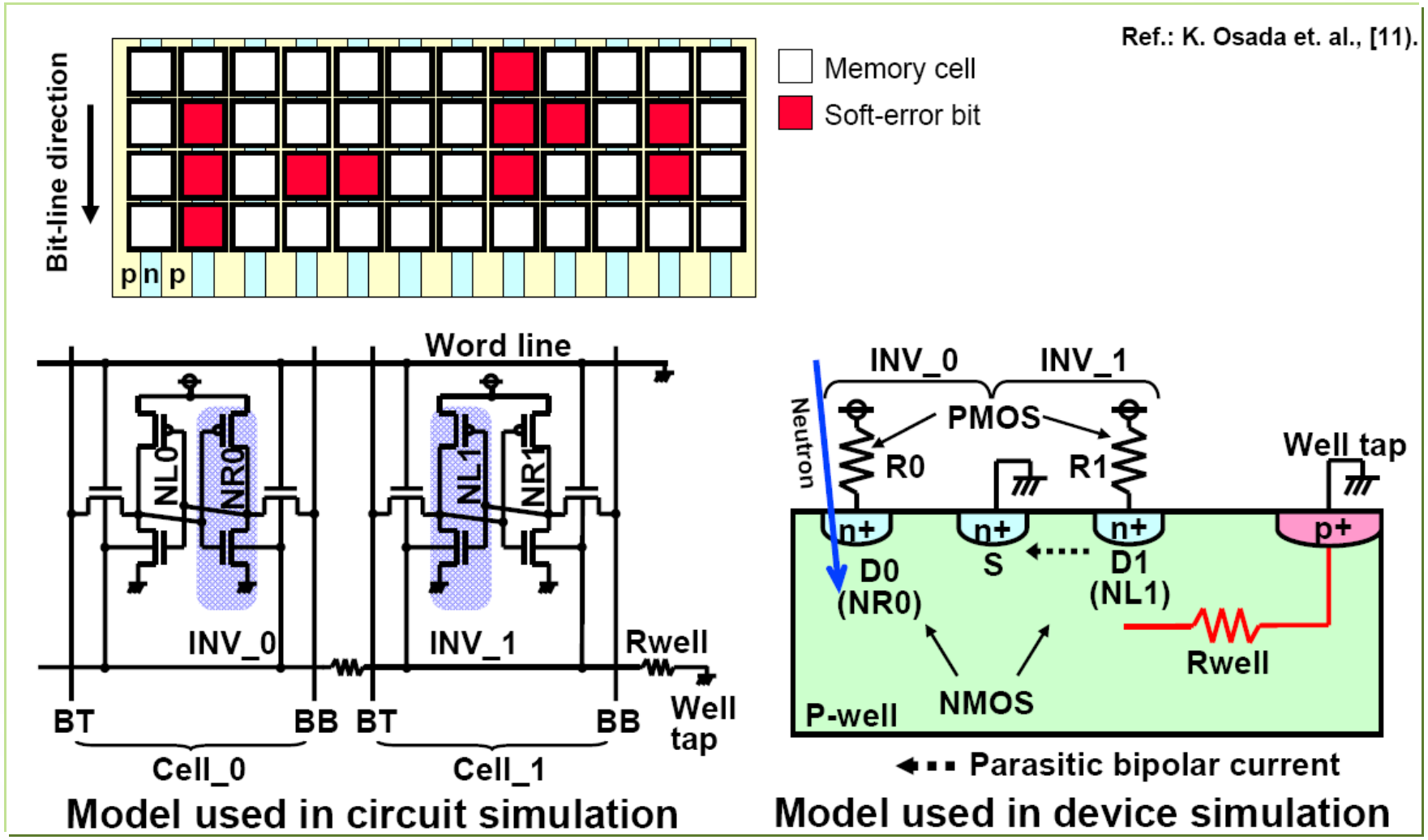
- Single Error Correction Double Error Detection (SECDED)
  - Hamming codes with additional parity



Wikipedia

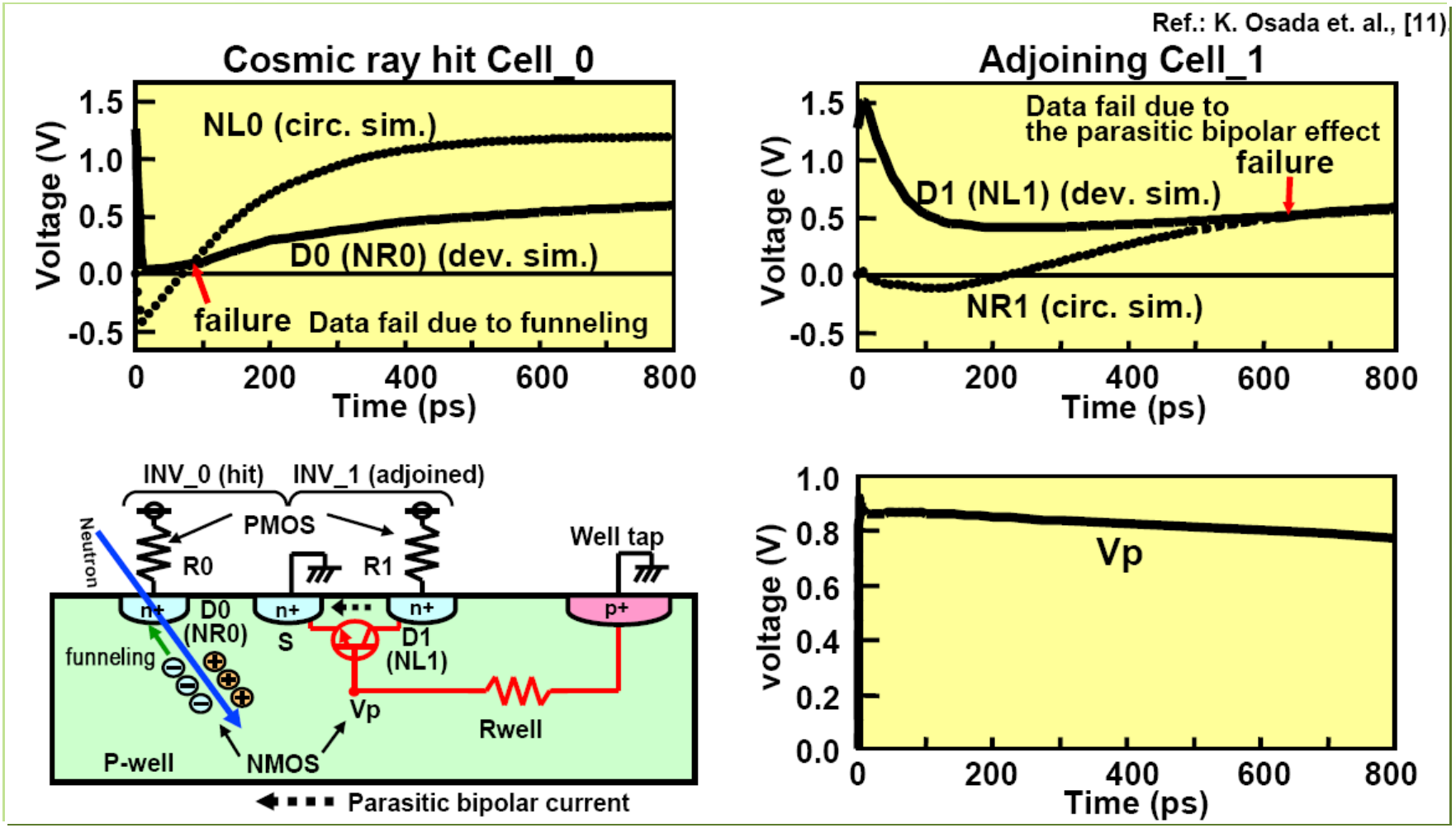
- Double Error Correction Triple Error Detection (DECTED)
  - BCH codes – higher decoding complexity

Multi-bit Errors



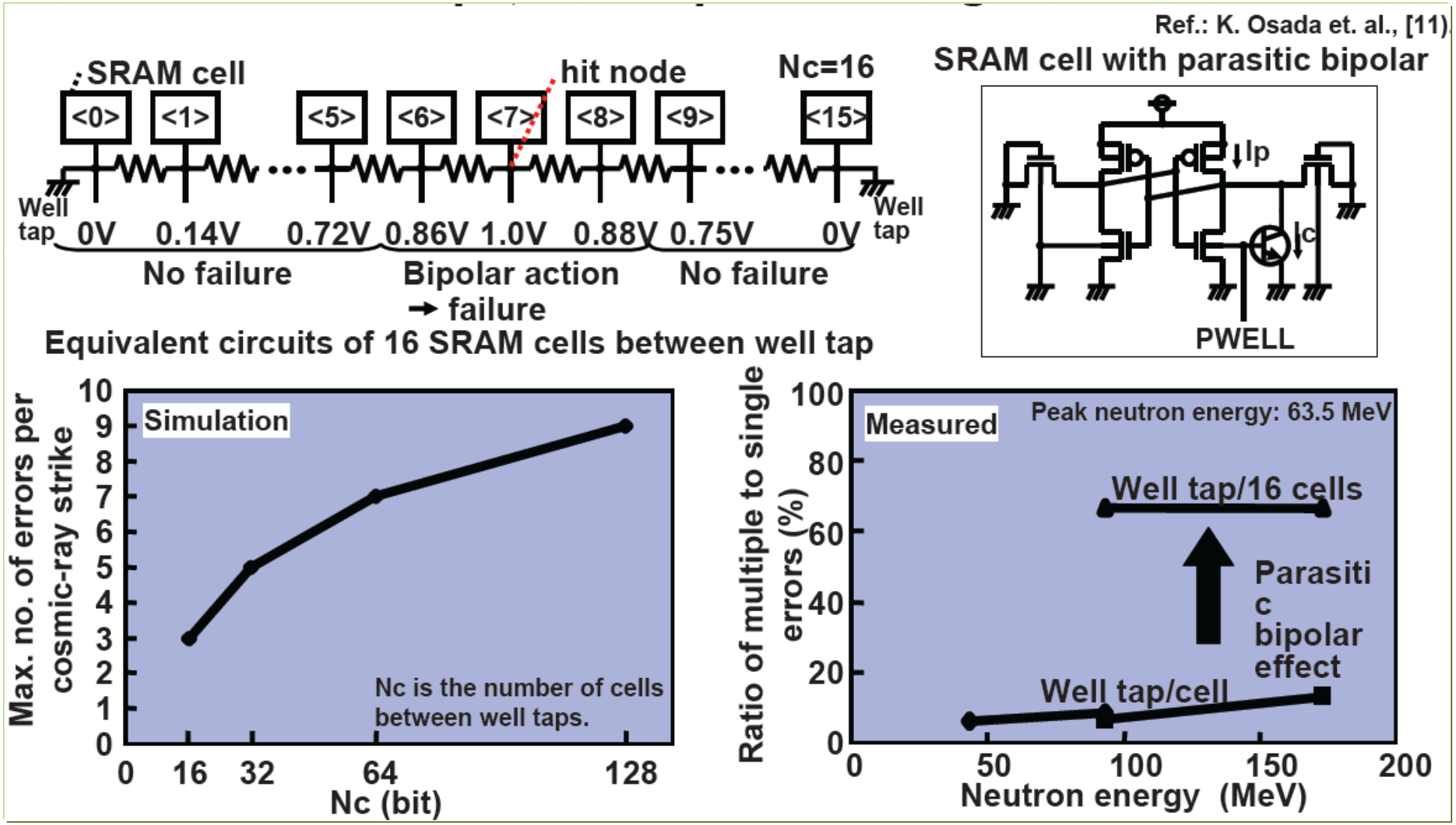
Kawahara, ISSCC'07 tutorial

Multi-bit Errors

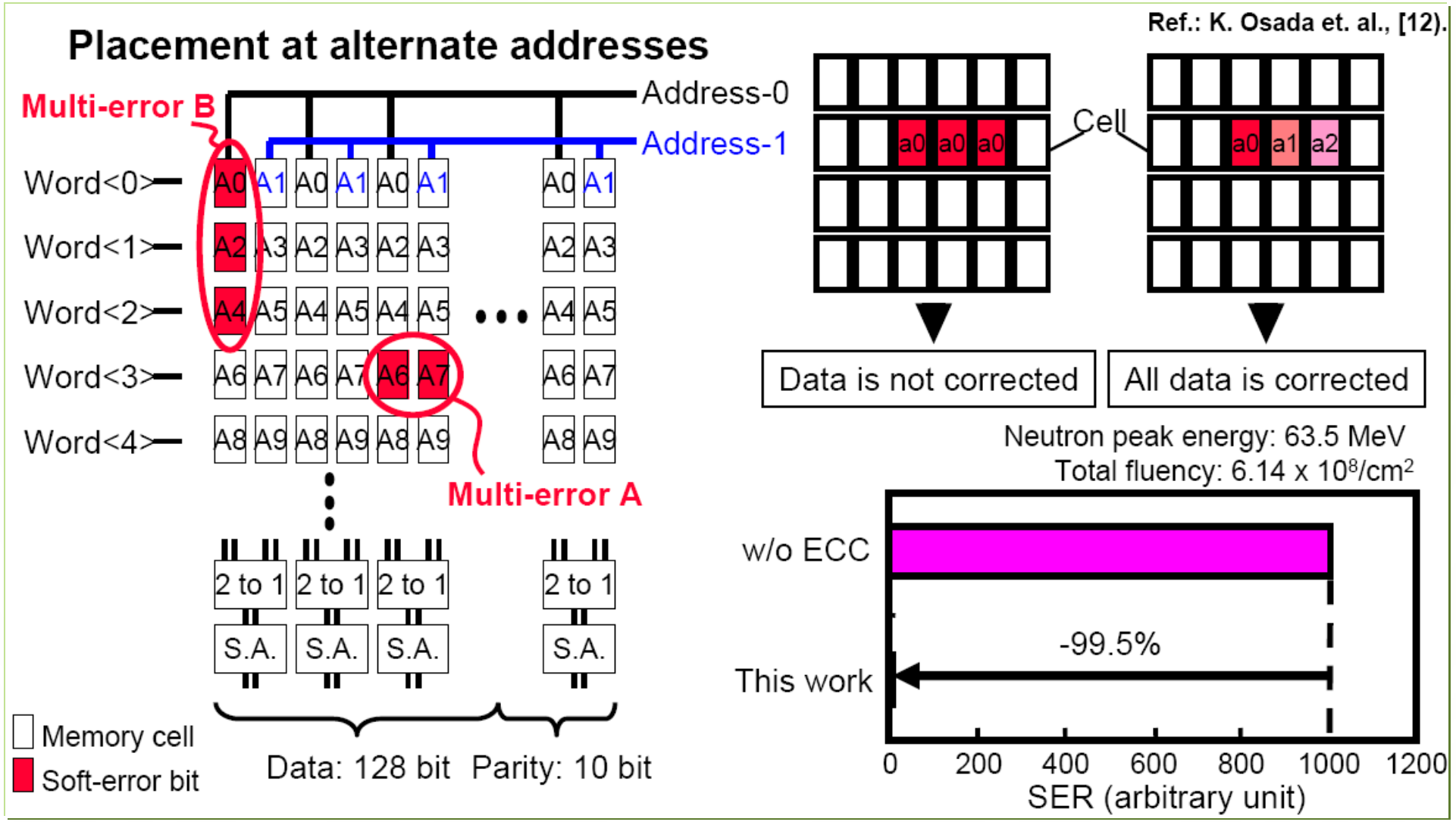


Kawahara, ISSCC'07 tutorial

Multi-bit Errors

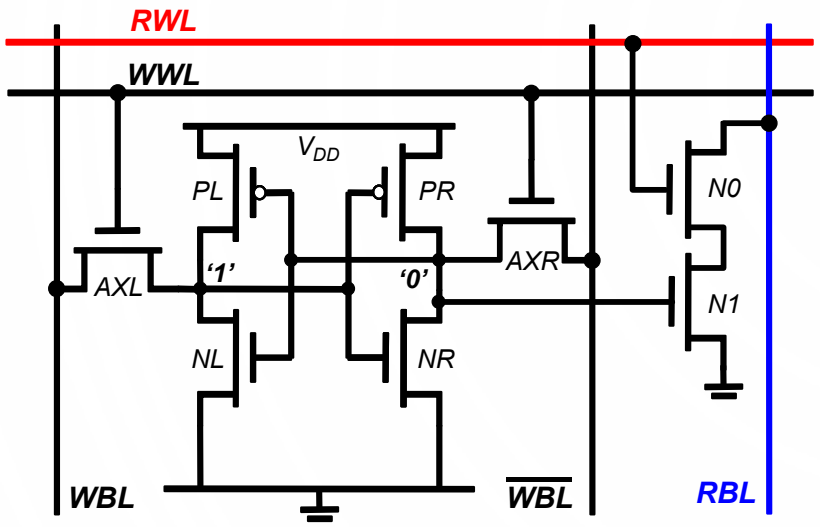


Multi-bit Errors: Interleaving



6T SRAM Alternatives

8T-SRAM



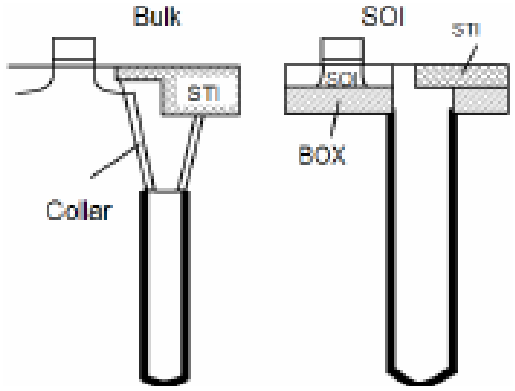
- Read circuit?
- Interleaving?

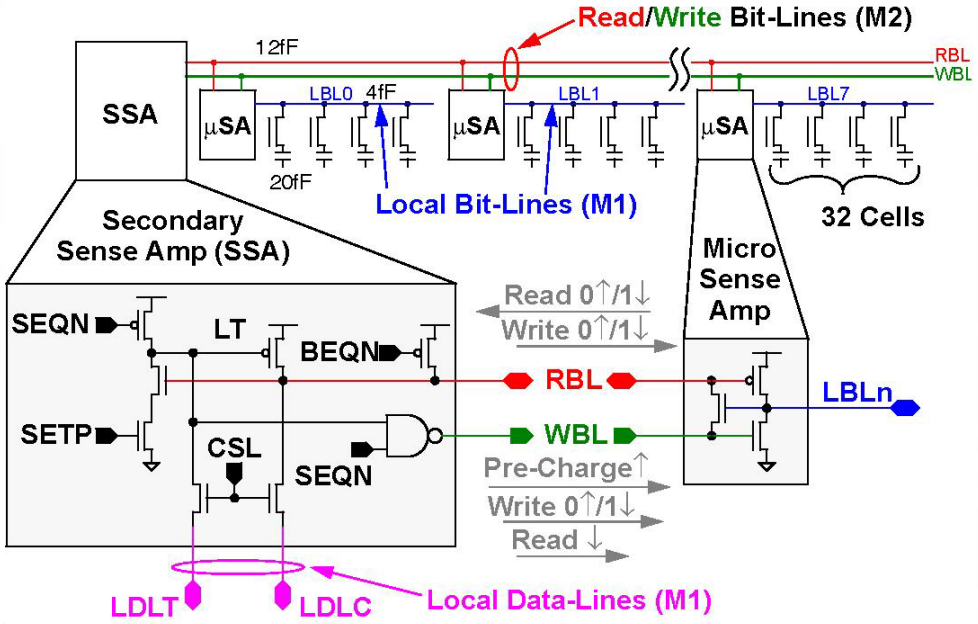
- Dual-port read/write capability (register-file-like cells)
- N0, N1 separates read and write
  - No Read SNM constraint
  - Half-selected cells still undergo read
- Stacked transistors reduce leakage

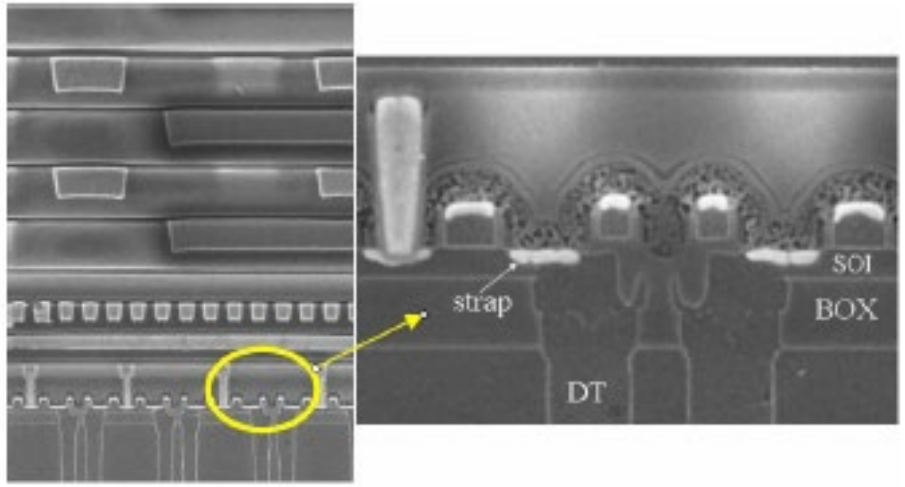
L. Chang, *VLSI Circuits* 2005

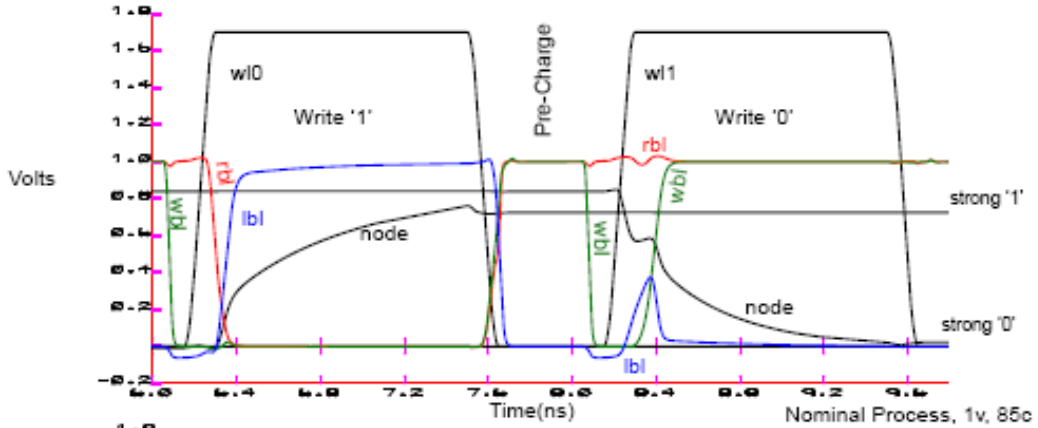
eDRAM

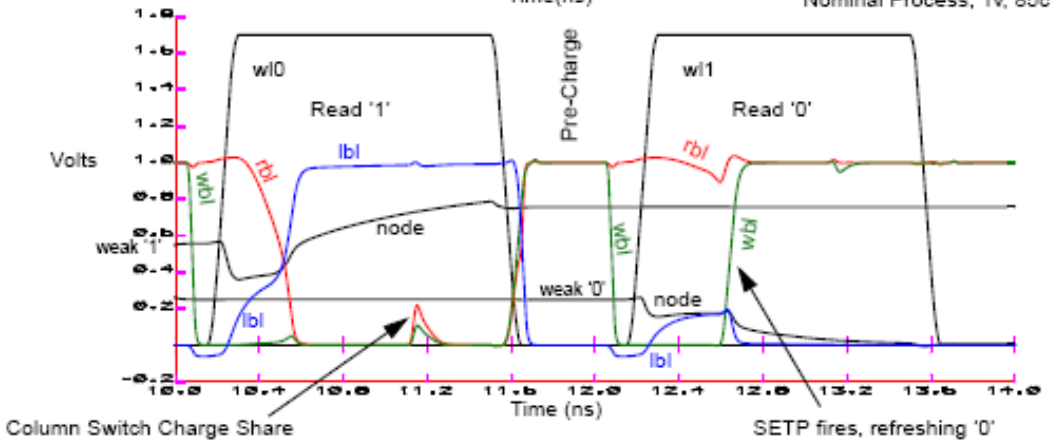
- Process cost: Added trench capacitor











Barth, ISSCC'07, Wang, IEDM'06

Crosspoint Memories

- Barrett, IRE Trans. Comp. 1961.

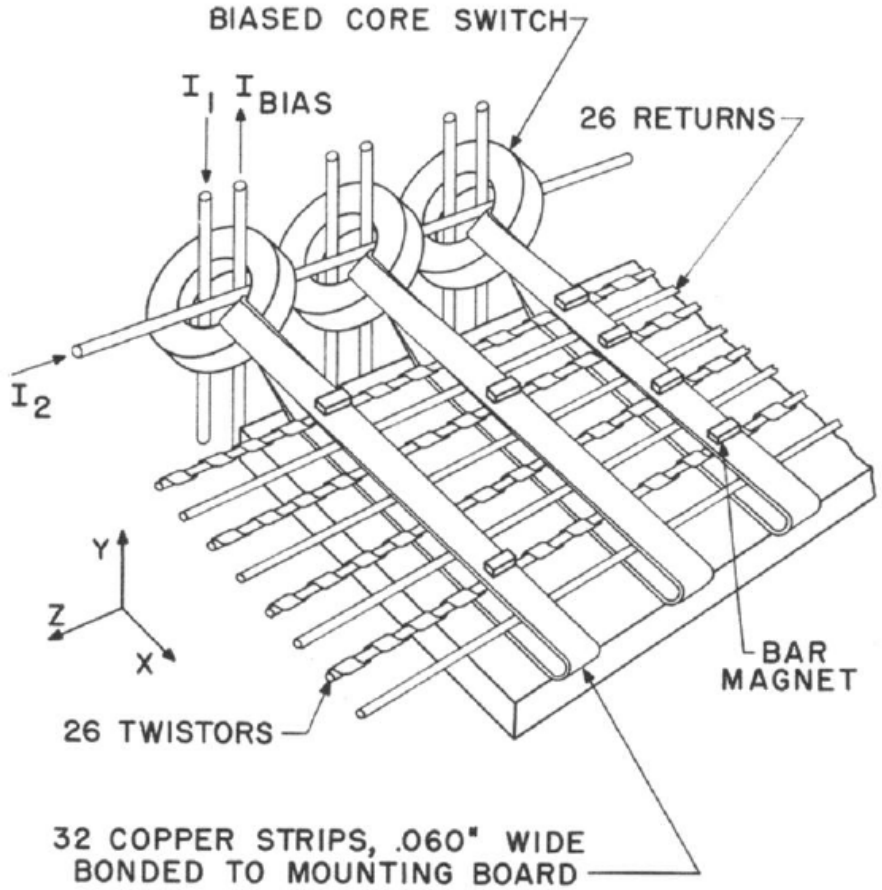
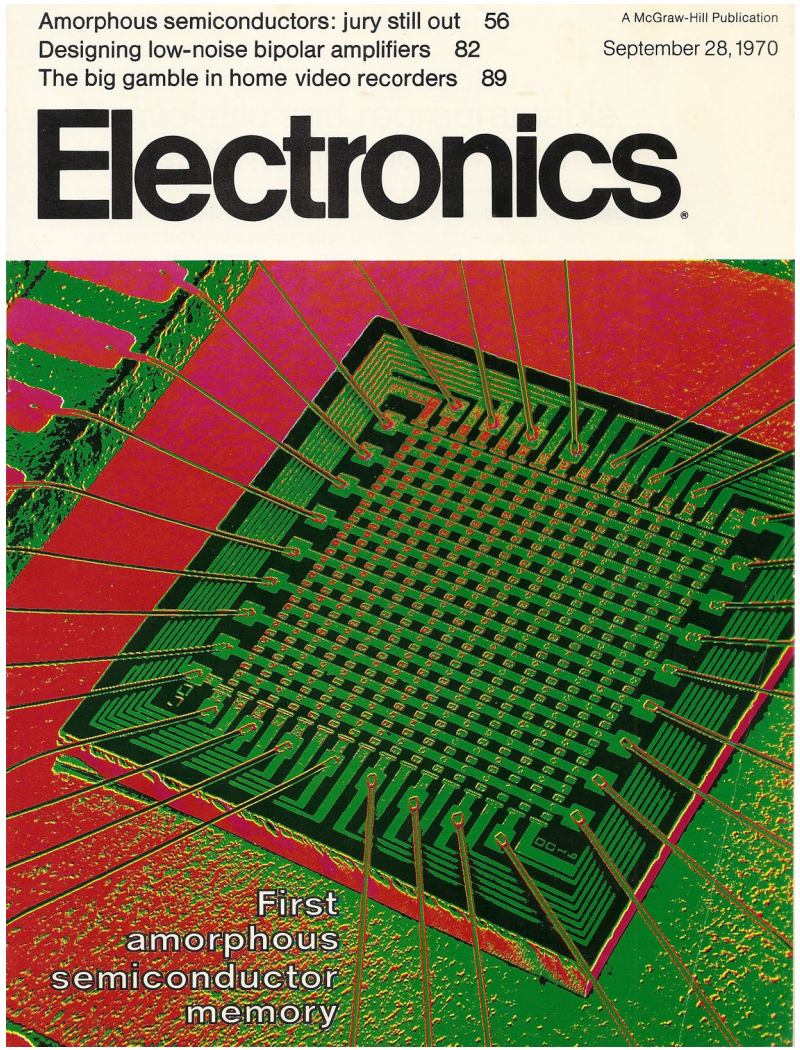
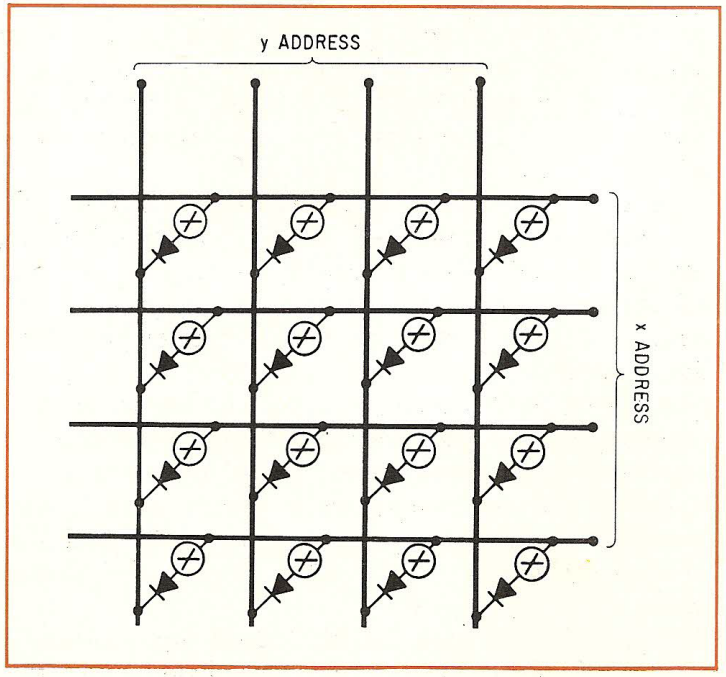


Fig. 2—Memory structure. I_1 and I_2 are access drive currents to core-selection switch. Presence or absence of a magnet over a twistor-strip solenoid crosspoint yields a "zero" or "one." Signals observed between twistor and return wire.

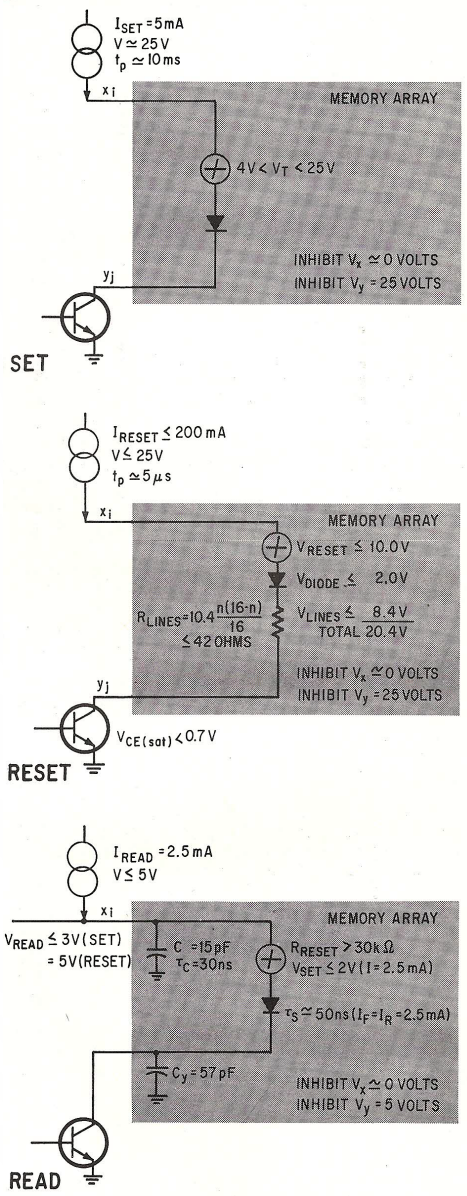
Crosspoint Memories



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

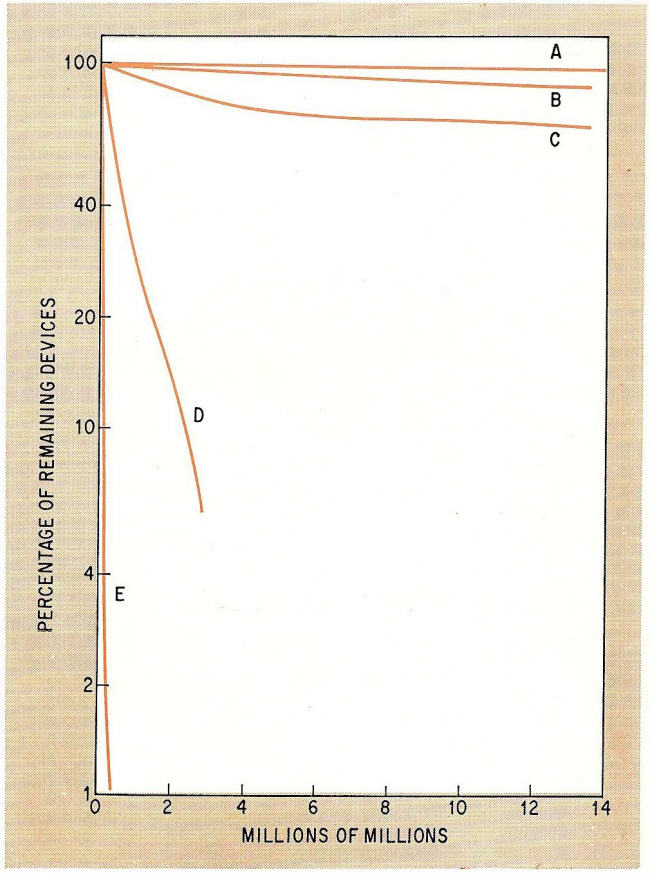


Crosspoint Memory



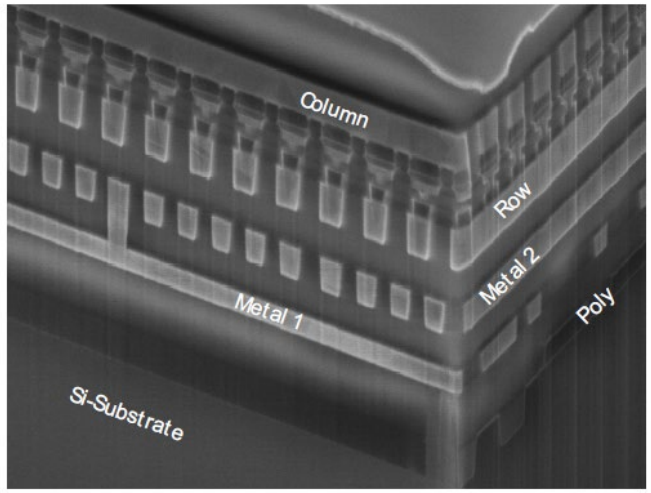
- Four modes
  - Form
  - Set
  - Reset
  - Read

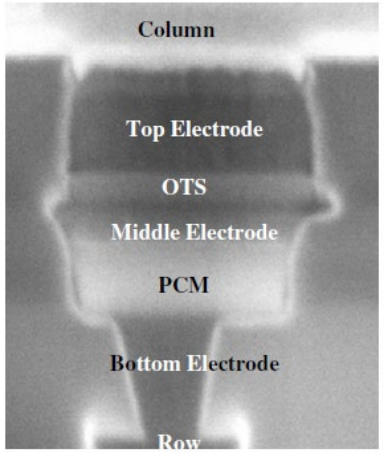
Endurance

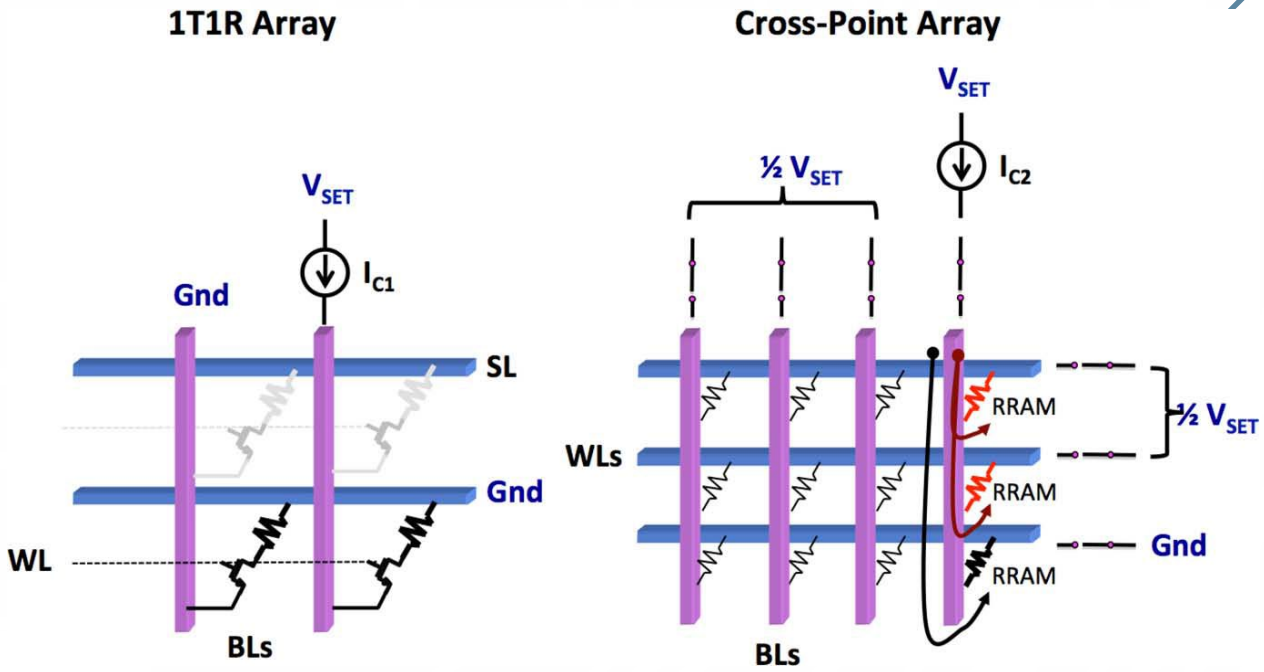


3D Crosspoint Arrays

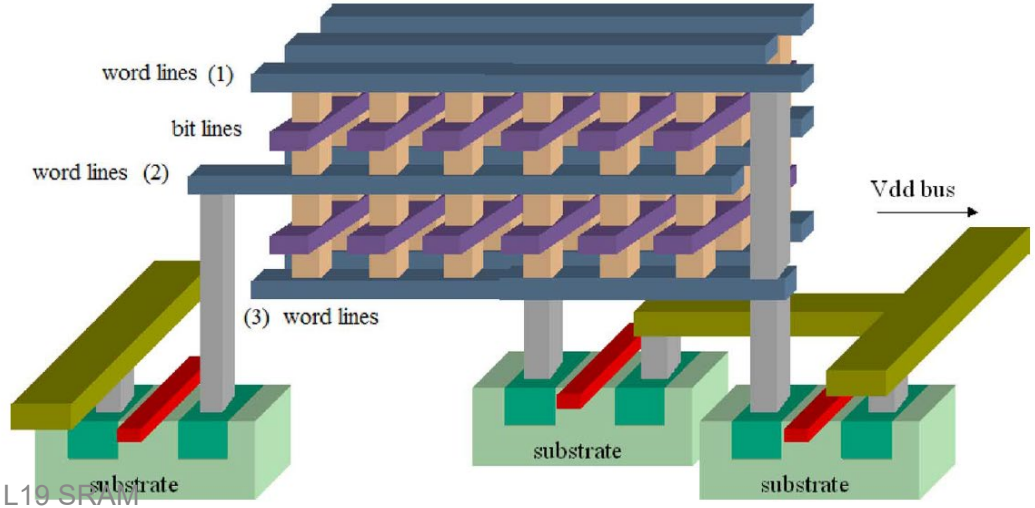
- Kau, IEDM'09







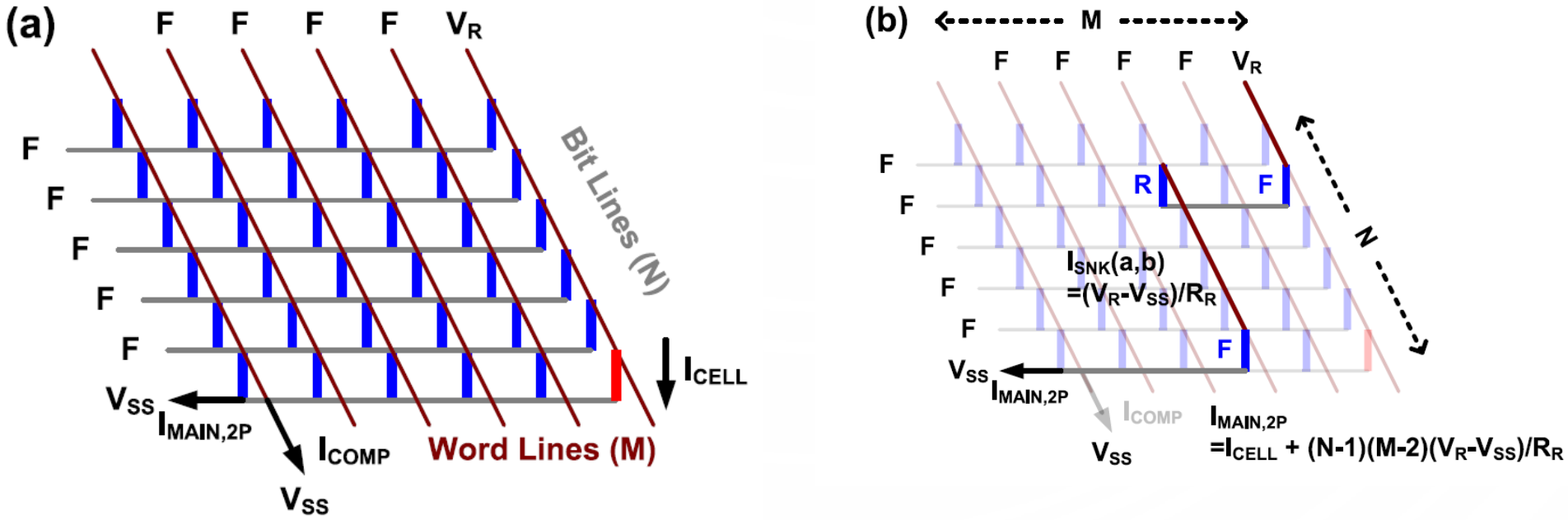
- Yeh, JSSC'15



- Ou, JSSC'11

Crosspoint Arrays

- Read and sneak currents



Bae, TED 4/17

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

- Spring break
- Low power design