EE-382M VLSI–II
SRAM Circuit Design

Spring 2017

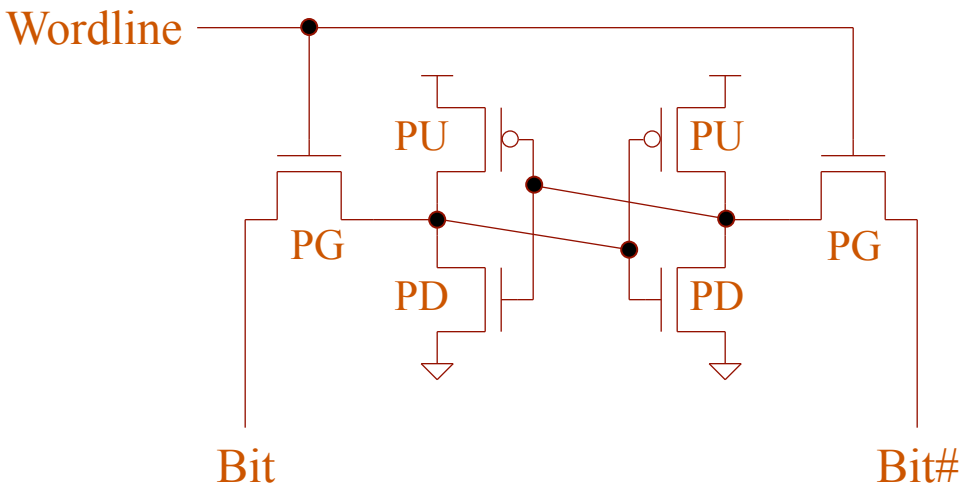
Mark McDermott
Gian Gerosa
Steve Sullivan

Class Agenda

- SRAM Overview
- Read/Write operation
- Decoders/Word-line drivers
- Bit-cell Design
- Static Noise margin / stability
- 6T bit-cell layout
- Alpha Particles & Soft Error Rate
- Charge Sharing
- Bit-line Circuits
- Write Circuits
- Sense Amplifiers
- Putting it all together (including redundancy)

The Basic 6T BIT-CELL

The most common SRAM memory cell uses 6 transistors; the same bit-lines are used for reads and writes.

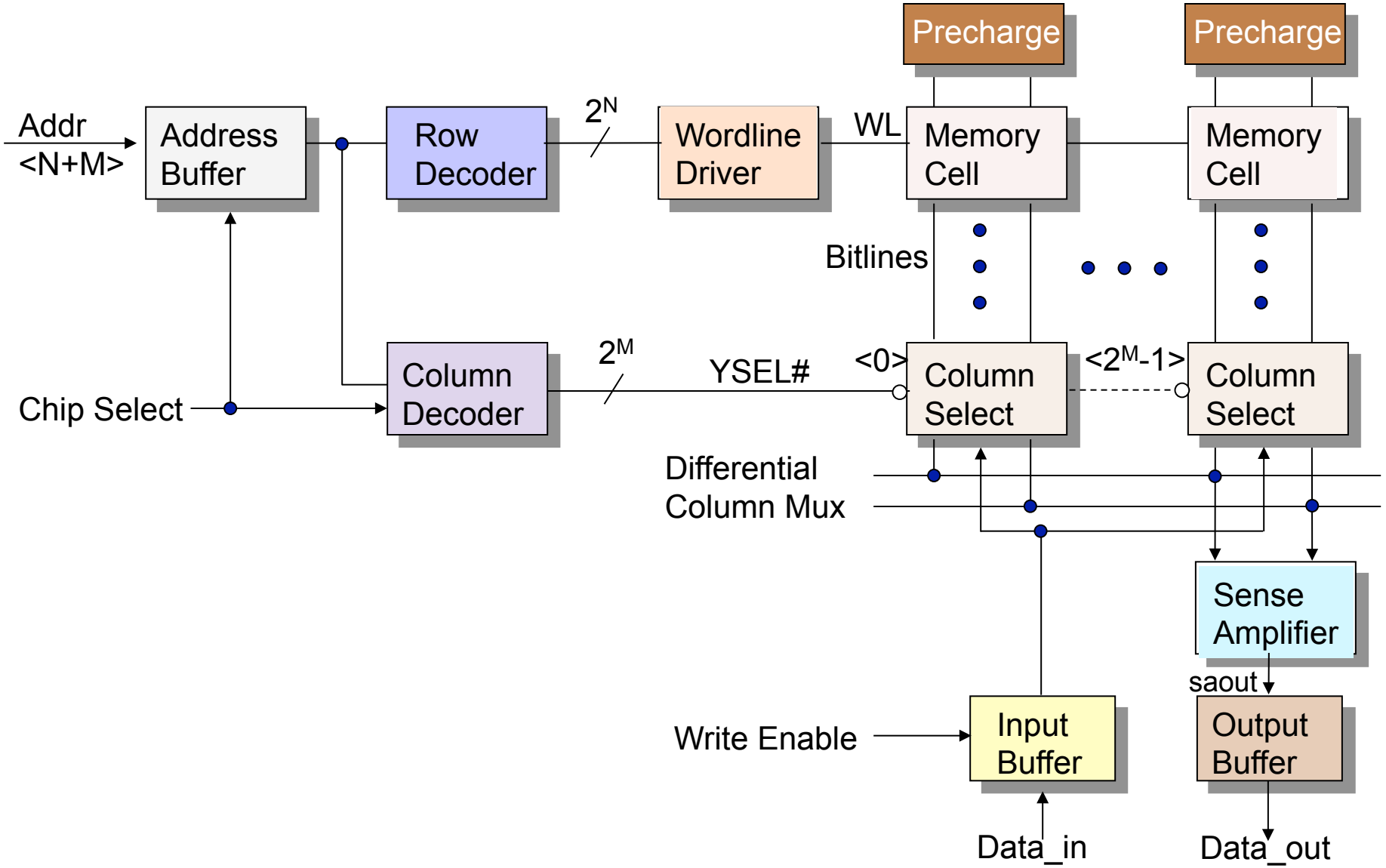


PG: pass gate
PU: pull-up
PD: pull-down

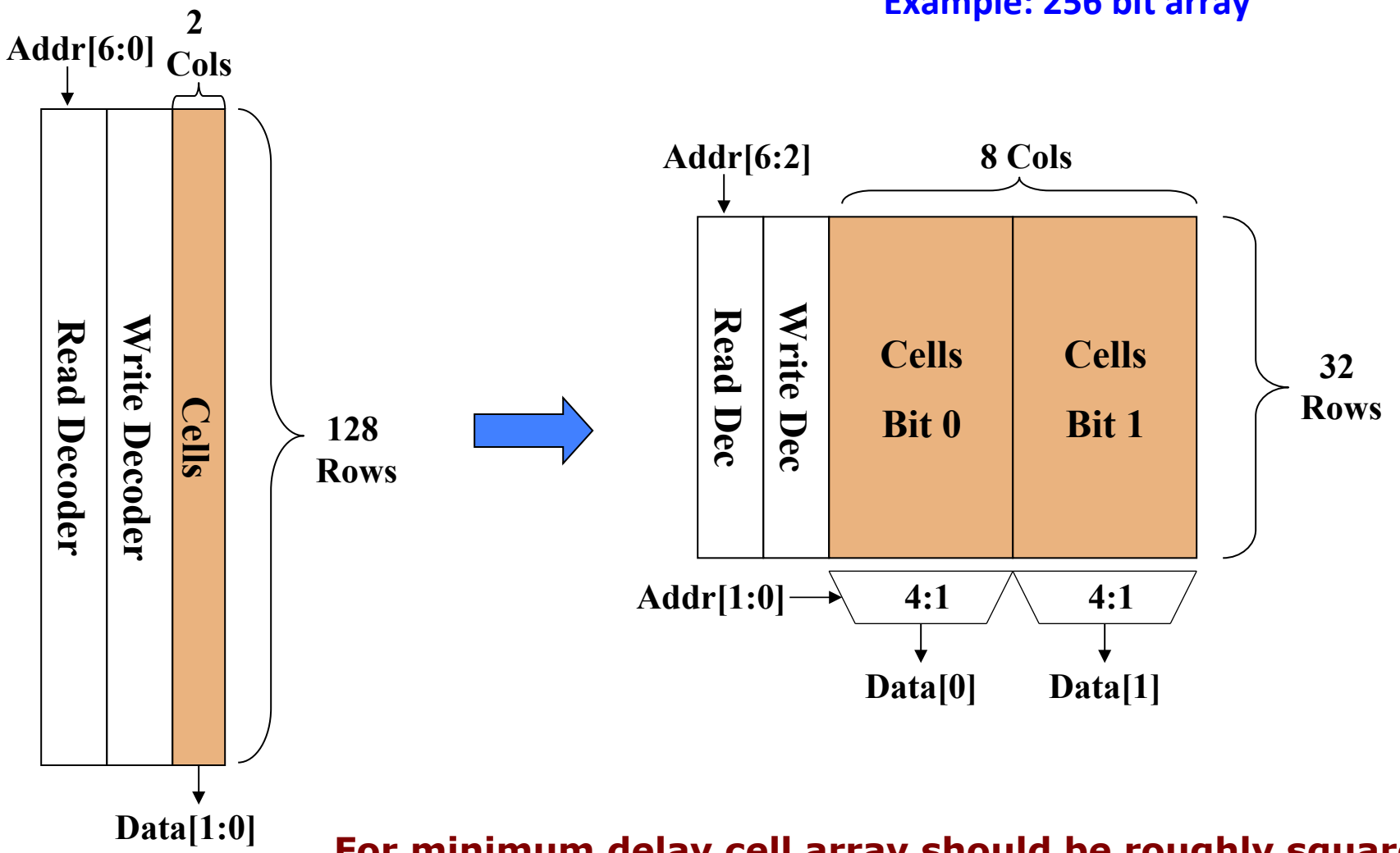
PD > PG for read stability
PG > PU to enable writes

Can be made with standard logic manufacturing process.

SRAM Block Diagram

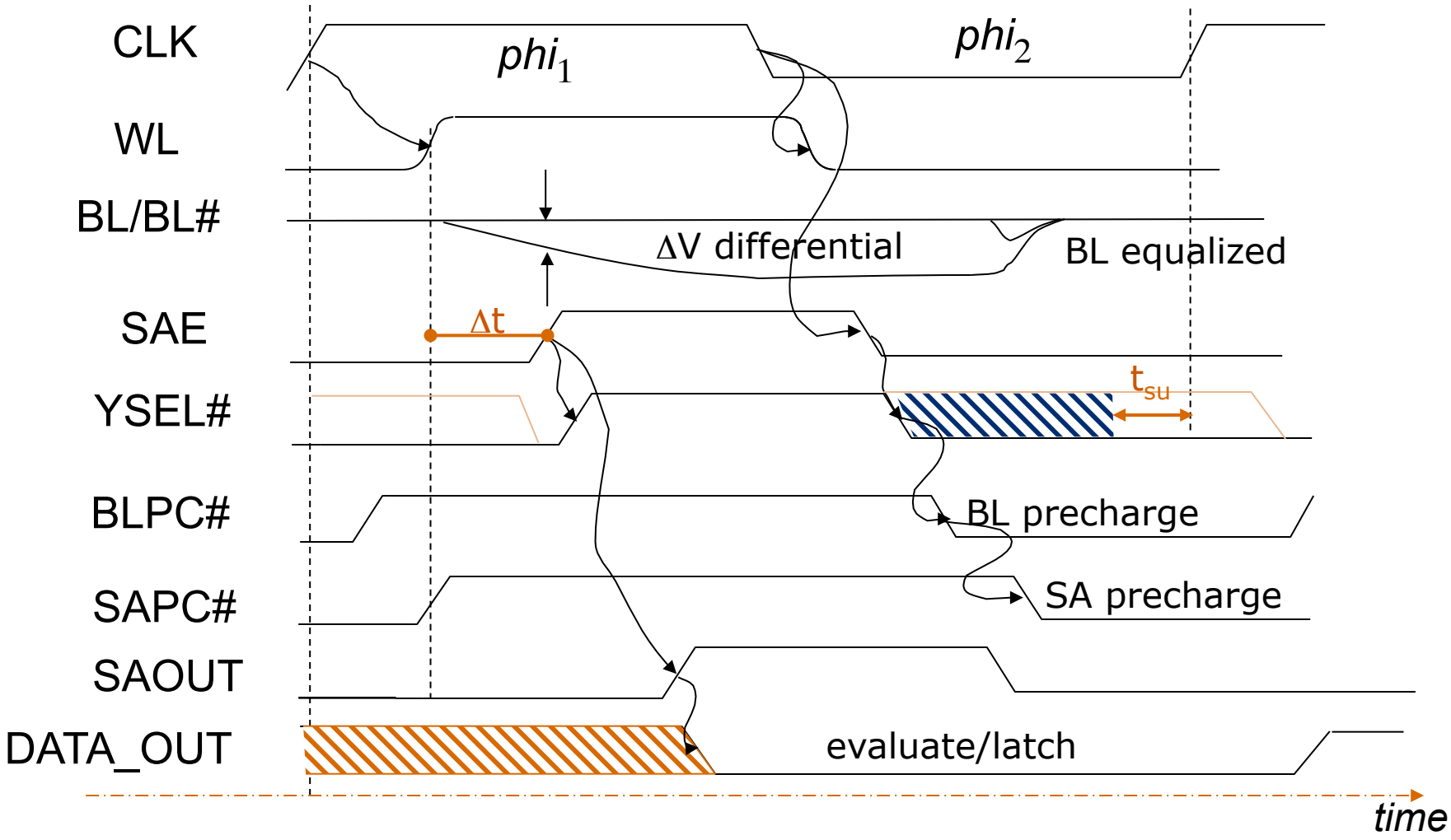


ARRAY ORGANIZATION



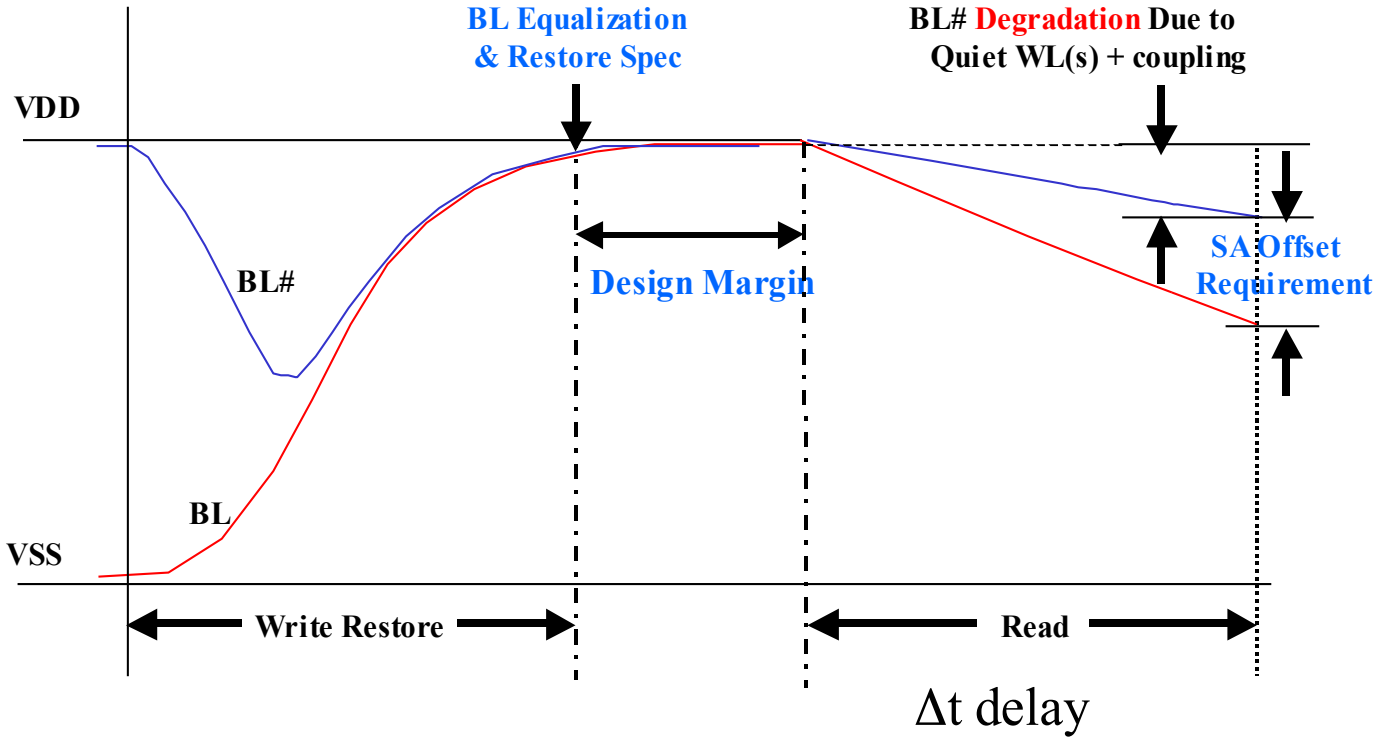
For minimum delay cell array should be roughly square, if the bitcell row/column delays are equivalent.

READ TIMING


CLK
WL
BL/BL#
SAE
YSEL#
BLPC#
SAPC#
SAOUT
DATA_OUT
phi1
phi2
ΔV differential
BL equalized
Δt
tsu
BL precharge
SA precharge
evaluate/latch
time

READ REQUIREMENTS

- Pre-charge & equalize bit-lines from previous cycle
- Minimum "Design Margin" before next READ begins
- Delay requirement to allow sufficient bit-line voltage development



- NOTE: The Δt delay can be generated by a chain of inverter delays or by replica "dummy" row and column composed of bitcells.

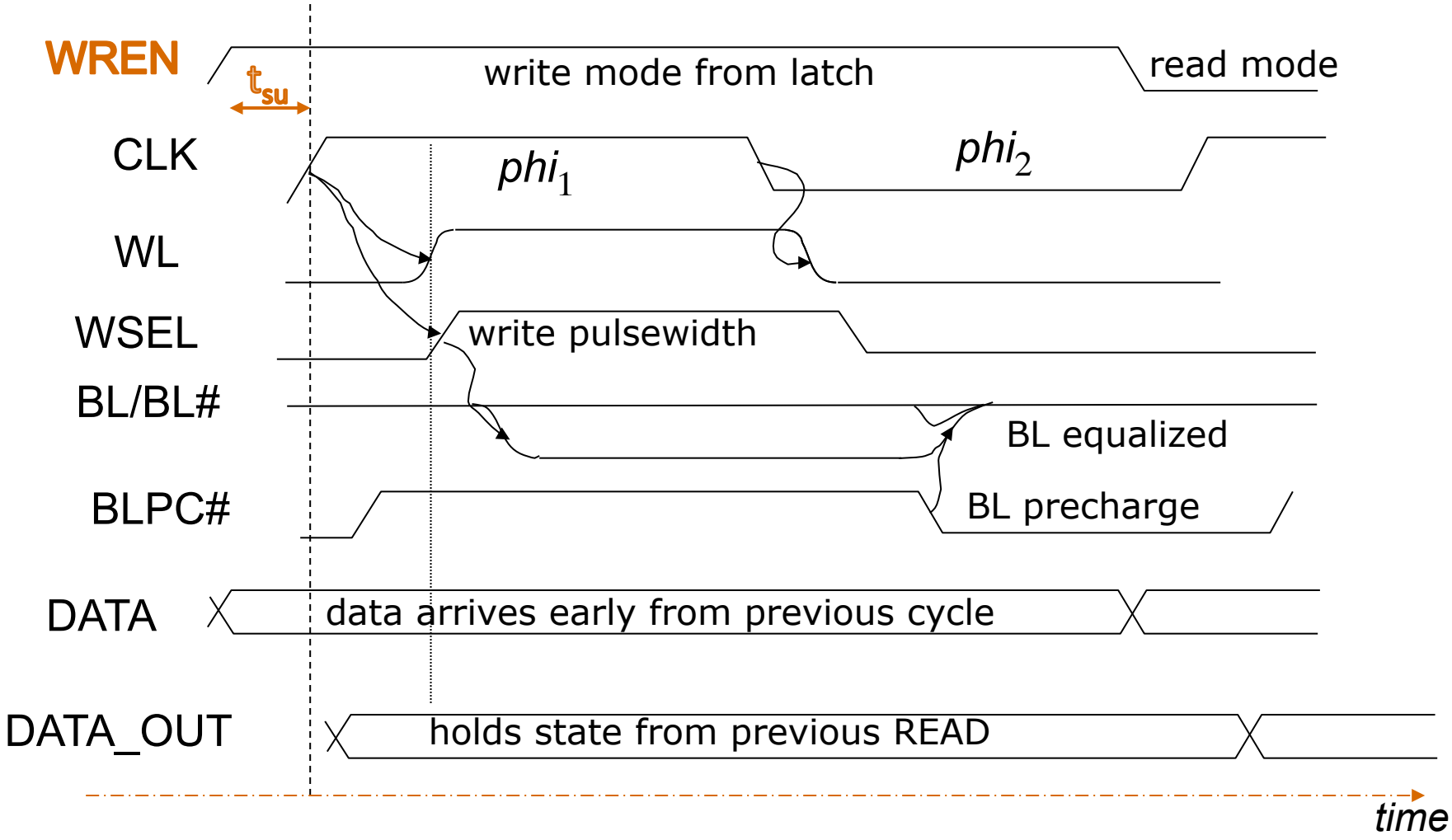
Basic READ Operation (part 1)

- Bit-lines are pre-charged to high potential & equalized
- Clock transitions HIGH to begin READ cycle (read_enable asserted)
  - Pre-charge shuts off & address is pre-decoded using NAND gates
  - Read mode enables word-line drivers (initially all word-lines "OFF")
- 2nd level of decoding performed by word-line driver (i.e, dynamic NAND/INV)
  - One word-line driver is fully decoded to assert word-line
- One of many rows drives many bit-lines BIT/BIT# through the weak bit-cells
- The voltages on the bit-lines moves very slowly since bit-line capacitance is large (wire plus diffusion capacitance of many pass gates)

Basic READ Operation (part 2)

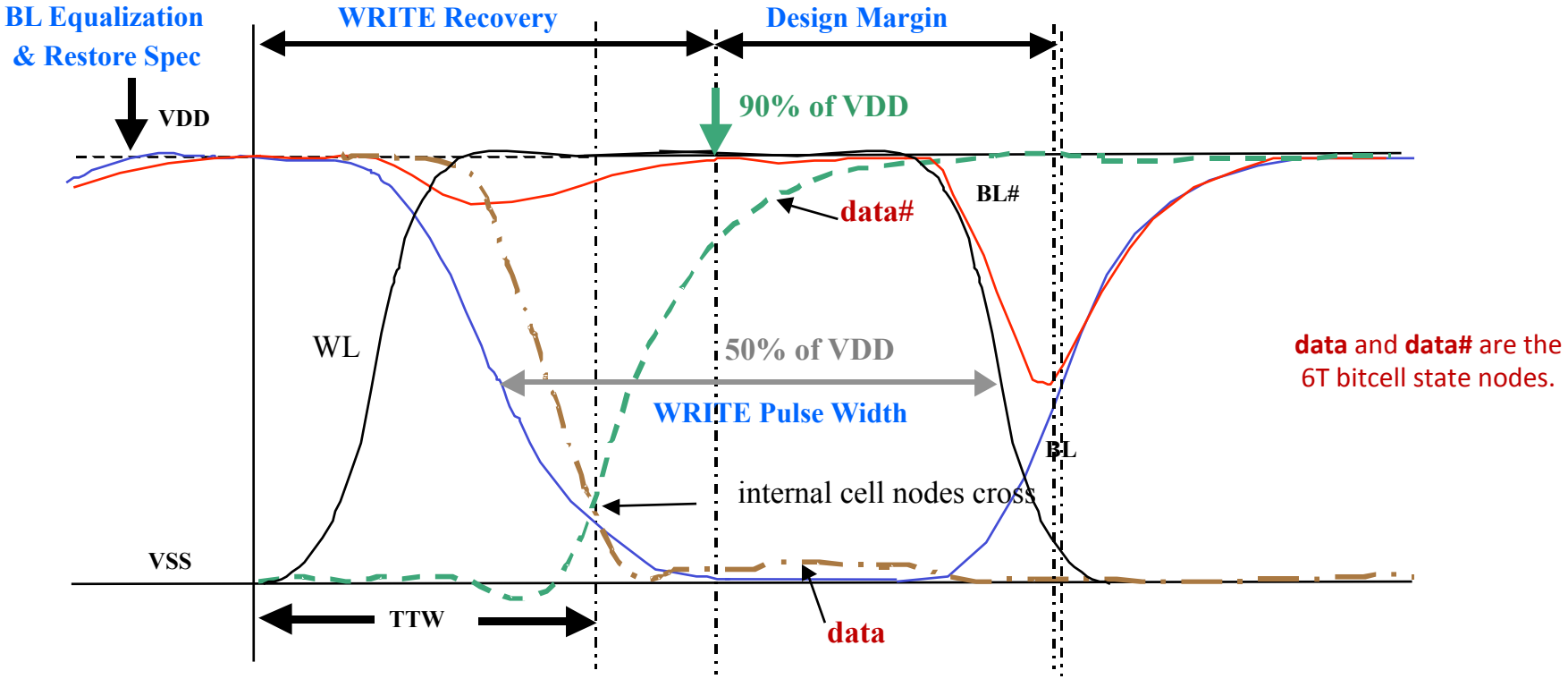
- 6-T bit-cell layout is as small as possible to allow a predetermined minimum voltage differential to develop on bit-lines BIT/BIT# within cycle time constraints.
  - Could be as small as 80 to 100 millivolts
- Column decoder allows a pair of bit-lines to drive the sense amplifier
- DRAMs & SRAMs use differential sense amplifiers with good common mode rejection ratio and high gain
- The sense-amp converts the small differential signal to single-ended full-rail voltage levels
- The sense-amp output is captured in a latch or flip-flop and fed to the output buffer
- Pre-charge must be asserted to drive bit-lines to VDD & equalize prior to next cycle

WRITE TIMING



WRITE REQUIREMENTS

- Pre-charge & equalize bit-lines from previous cycle
- WRITE can begin as soon as word-line is available
- Must guarantee minimum write pulse-width, data valid time and write recovery; internal "high node" reaches say 90% of VDD



- NOTE: Write pulse-width margin increases with lower frequency

Basic WRITE Operation (part 1)

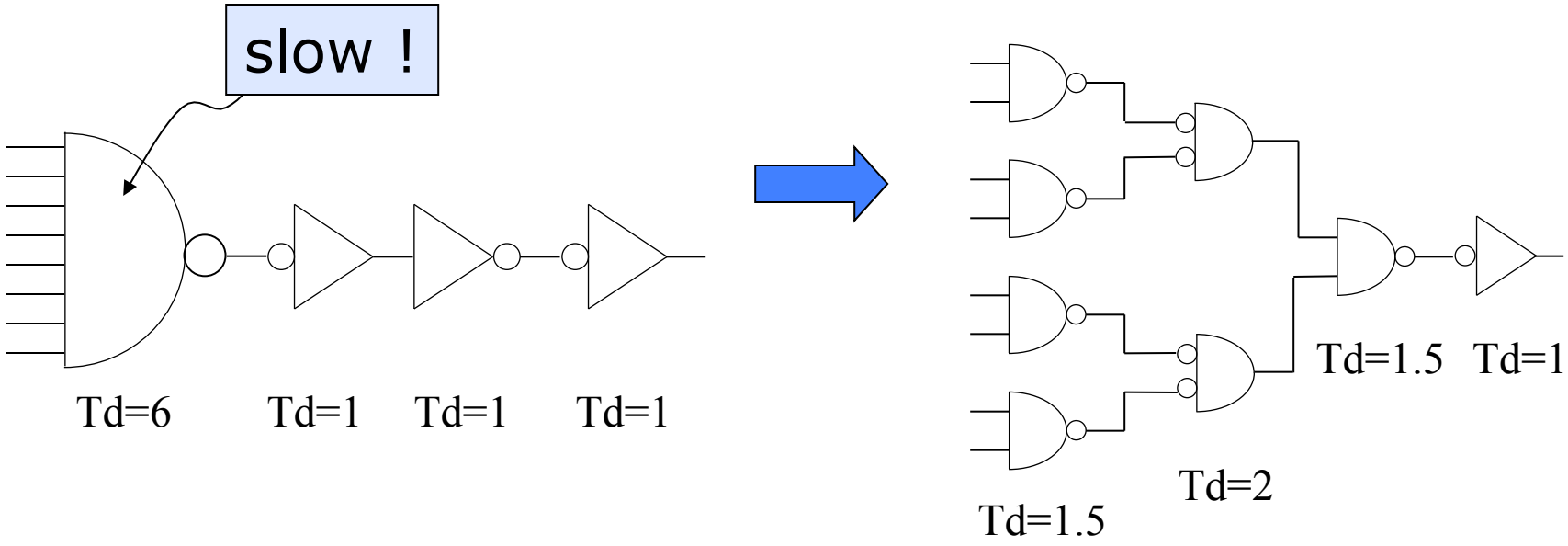
- Bit-lines are pre-charged to high potential & equalized
- Clock transitions HIGH to begin WRITE cycle (write_enable asserted)
  - Pre-charge shuts off & address is pre-decoded using NAND gates
  - WRITE mode also enables word-line drivers (bit-cells not driven by WRITE circuit normally perform a "pseudo READ")
- 2nd level of decoding performed by wordline driver (NAND/INV)
  - One word-line driver is fully decoded to assert word-line
  - One column mux is fully decoded to drive bit-line pair(s)
- data_in is usually setup early and pre-conditions the write circuit for writing a "1" or "0" into selected bit-cells
  - Write enable is often delayed to allow input data to settle out and prevent multiple writes
- One of many rows is selected for proper entry to be written
- Write driver is activated and quickly discharges either BL or BL#

Basic WRITE Operation (part 2)

- Cells are written by discharging the "high" side storage node through the weak NMOS pass gate of the bit-cell
- Other bit-lines are held to VDD-Vt (weak cross-coupled PMOS keepers helps drive opposite bit-lines towards high voltage)
- Normally, sense-amps are not enabled to save power
  - If data_out needs to be updated during the write, a bypass MUX is sometimes provided to send written value directly to output buffer
- Data for write must guarantee a hold time during write operation
- Pulse width must guarantee internal cell node completely recovers by the end of the write cycle for proper operation
- Finally, write circuit shuts OFF & PMOS devices activated to pre-charge bit-lines up to VDD

BASIC DECODERS

Logical decode is simply a wide AND gate

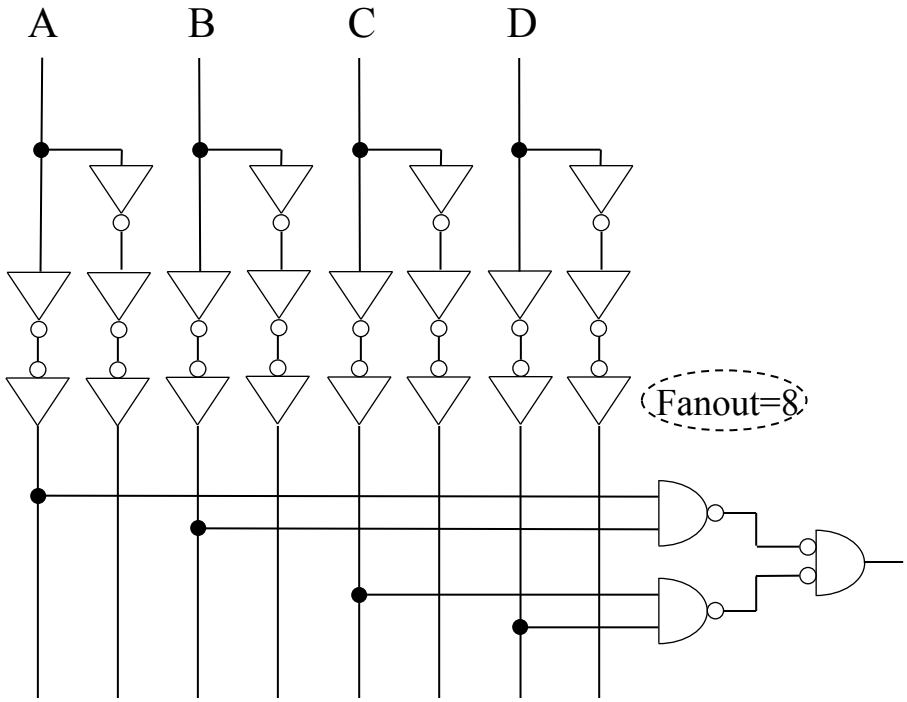


Total = 9 Inv Delays

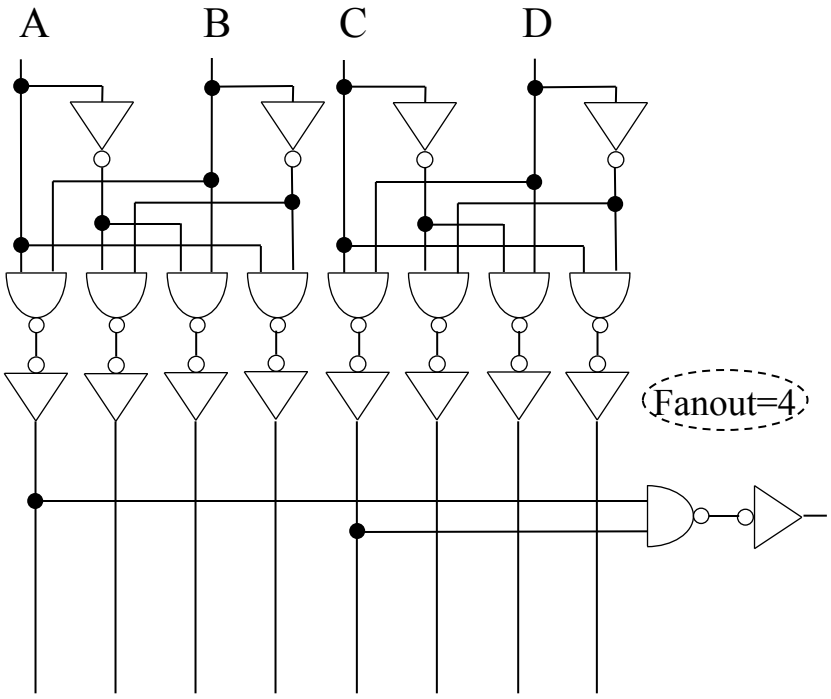
Total = 6 Inv Delays

Multiple stages of low fan-in gates are faster than a single large fan-in gate.

PRE-DECODING



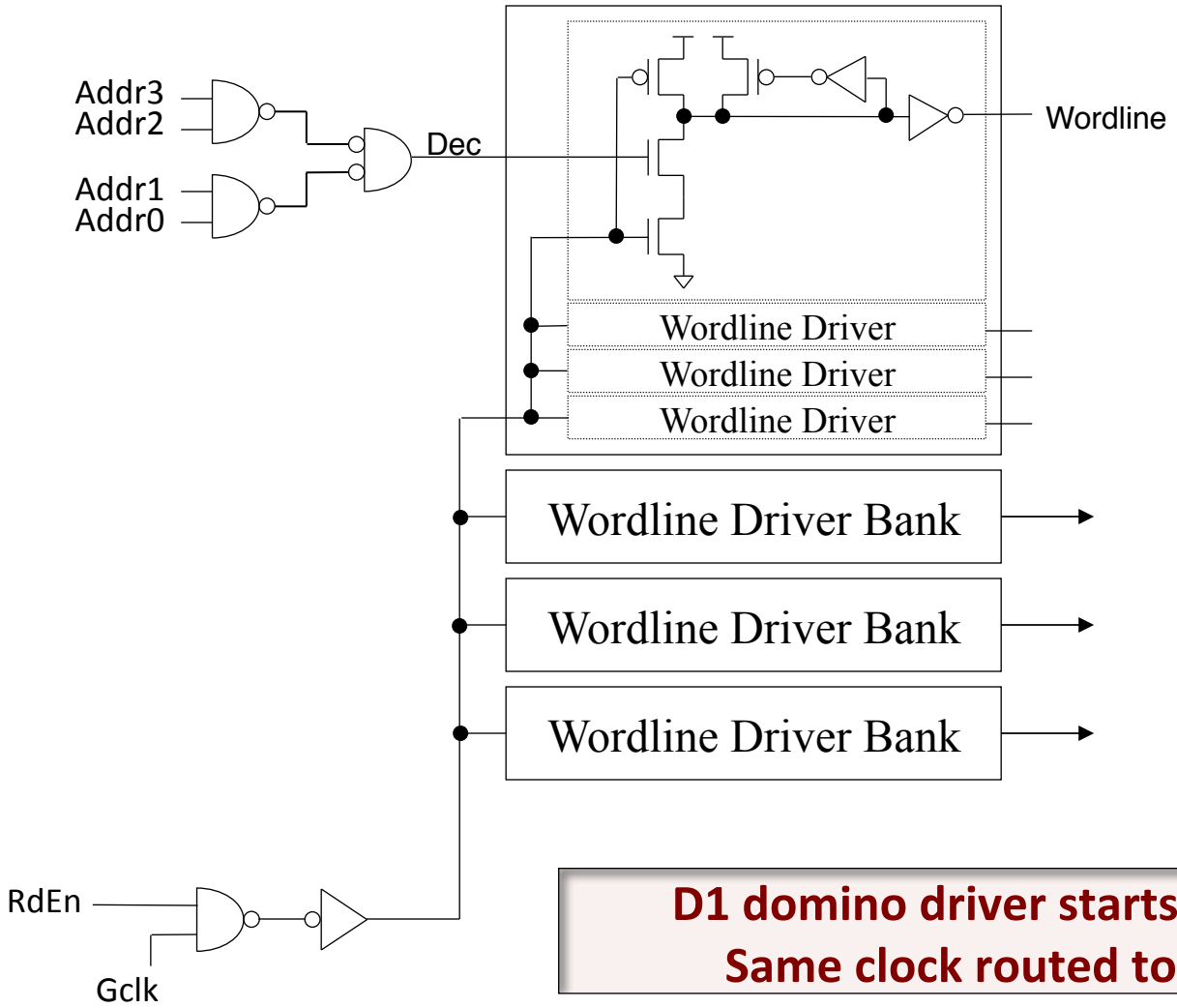
Each two bit pair requires 4 wires to route:
(A, A#, B, B#)



Each two bit pair requires 4 wires to route:
(AB, A#B#, A#B, AB#)

Predecode can reduce delay by cutting gate load on address lines.

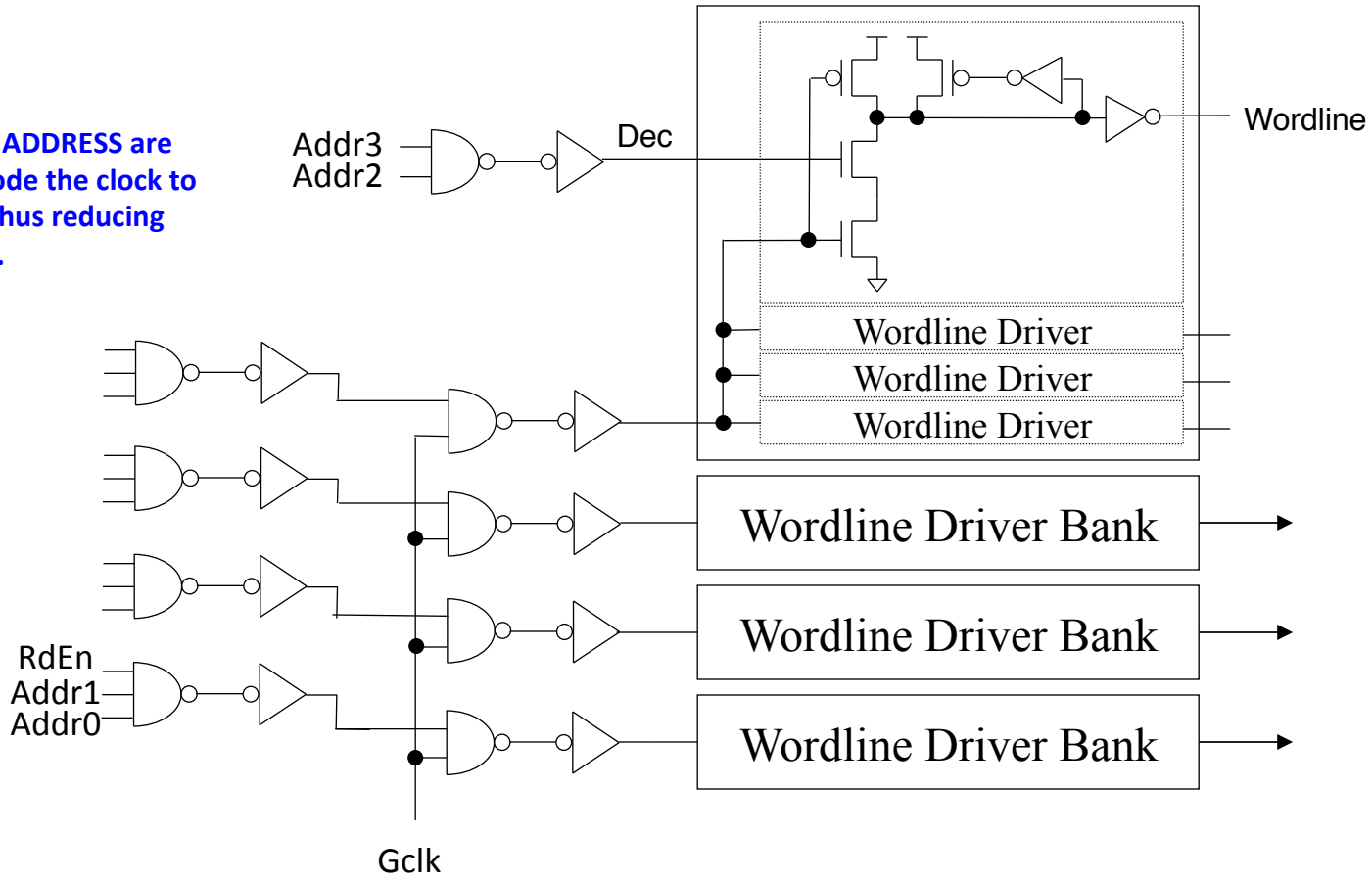
STANDARD WORD-LINE DRIVERS



D1 domino driver starts off wordline
Same clock routed to all drivers

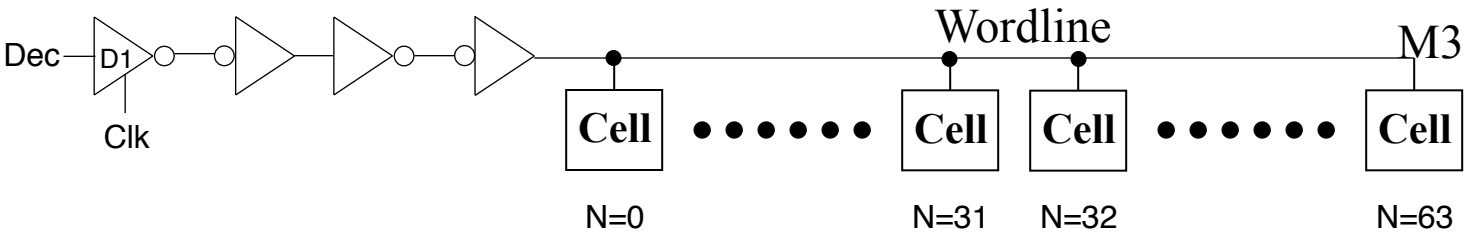
BANKED WORD-LINE DRIVERS

2 bits of the ADDRESS are used to decode the clock to each bank, thus reducing clock power.

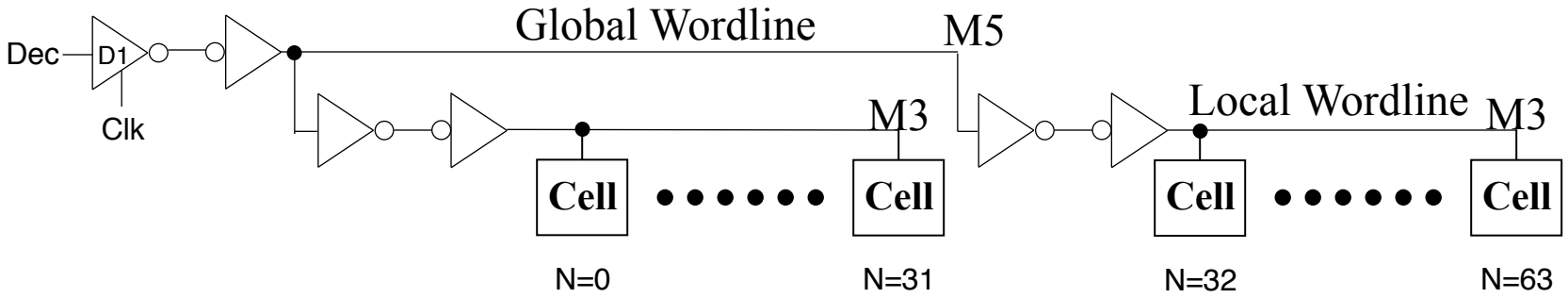


Enabling banks of wordline drivers can reduce decode delay and save power, however there is an area overhead

HIERARCHICAL WORD-LINES



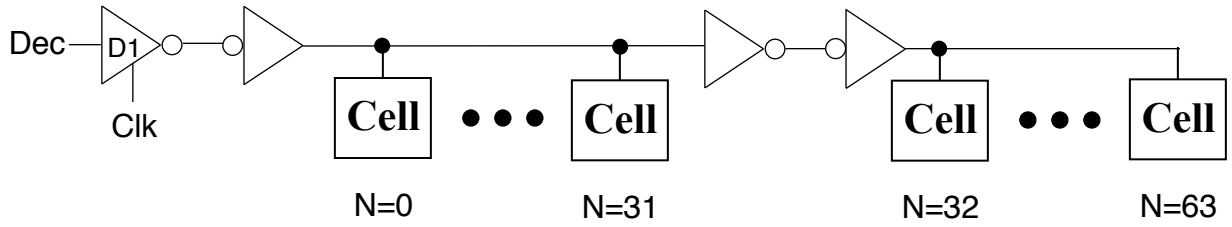
Large numbers of cells on the word-line will cause RC delay & noise problems.



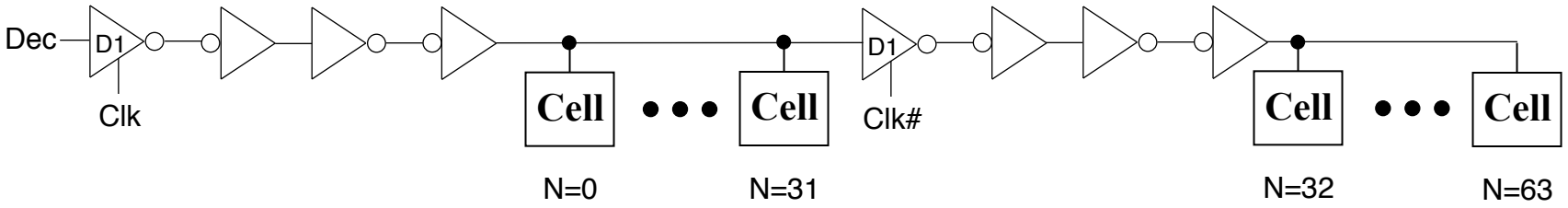
RC is reduced by breaking word-line into local segments at the cost of the use of an extra layer of metal or an extra routing track.

REPEATED WORD-LINES

If extra wires are not available, buffers or clocked repeaters may be necessary.



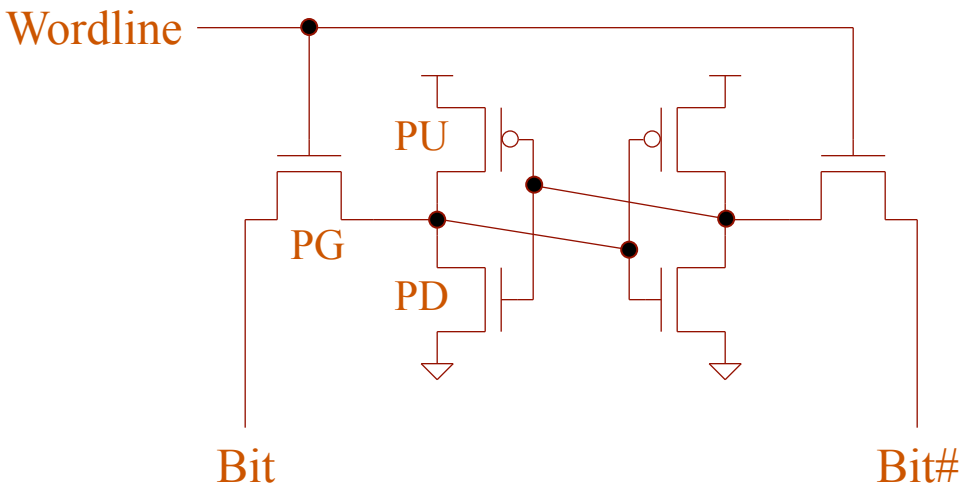
Buffered wordlines discharge bitlines later in the same phase.



Latched wordlines discharge bitlines in the next phase

The 6T BIT-CELL

The most common SRAM memory cell uses 6 transistors; the same bit-lines are used for reads and writes.

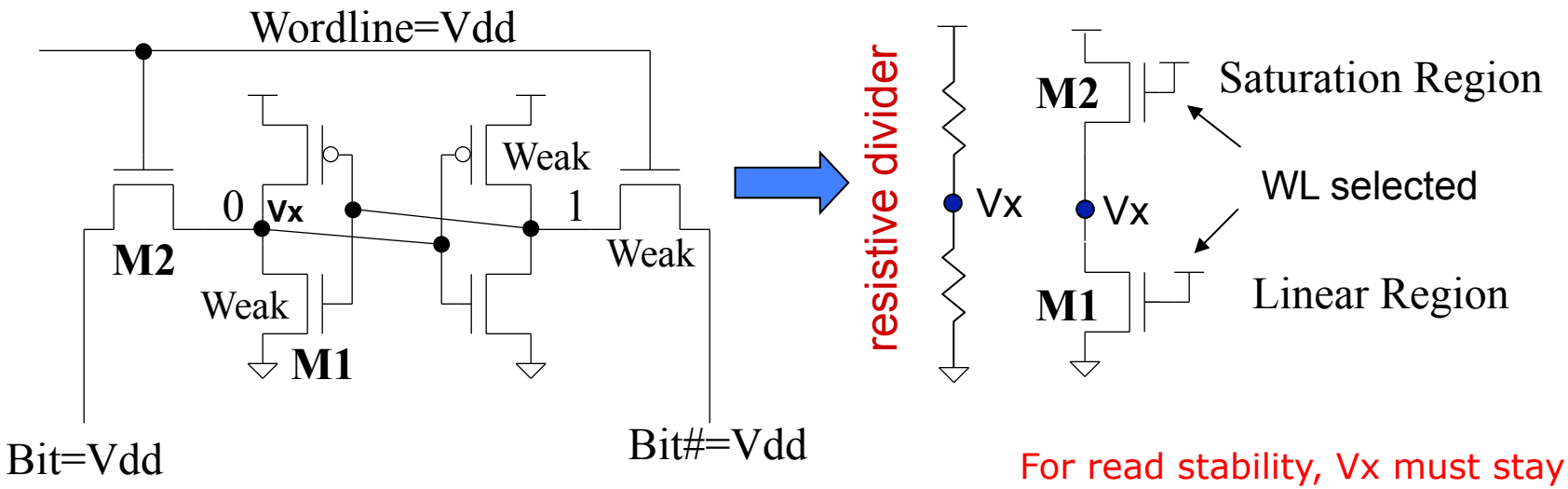


PG: pass gate
PU: pull-up
PD: pull-down

PD > PG for read stability
PG > PU to enable writes

Can be made with standard logic manufacturing process.

6T CELL – READ STABILITY



For read stability, Vx must stay below trip point of inverter.

I(Saturation) ∝ $(Vdd-Vt)^{\alpha}$ I(Linear) ∝ (Vdd-Vt)

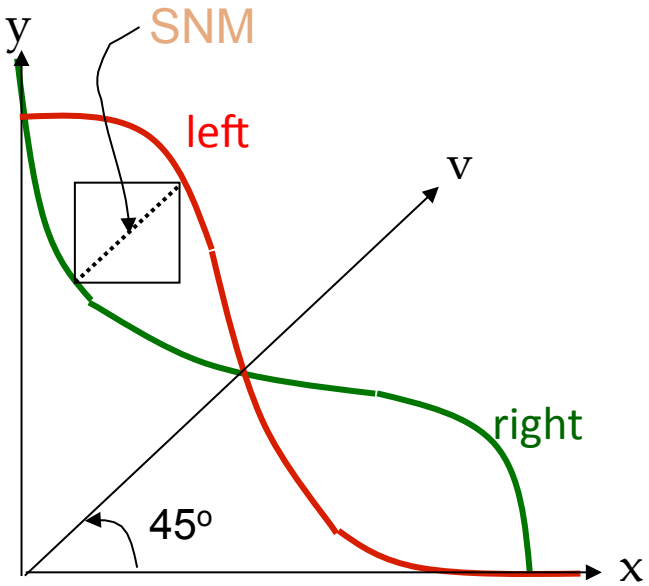
High voltage and high temp (which lowers Vt) will increase the current of M2 faster than M1 because M2 is in saturation.

This will cause the highest Vx voltage and be worst case for read stability.

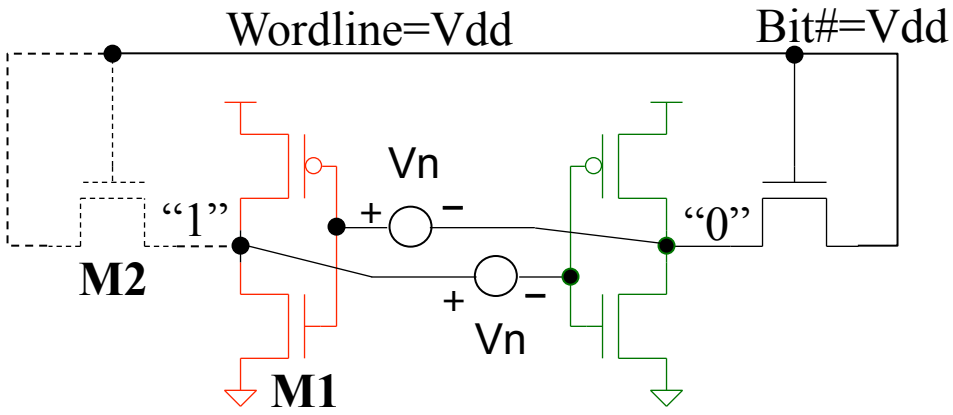
STATIC NOISE MARGIN [3]

- Insert "noise" source V_n in feedback path between inverters
- Generate inverter characteristic & mirror about 45° angle
- Keep access device to "low" side tied to VDD (worst case)
- Find diagonal of maximum square

diagonal of maximum square



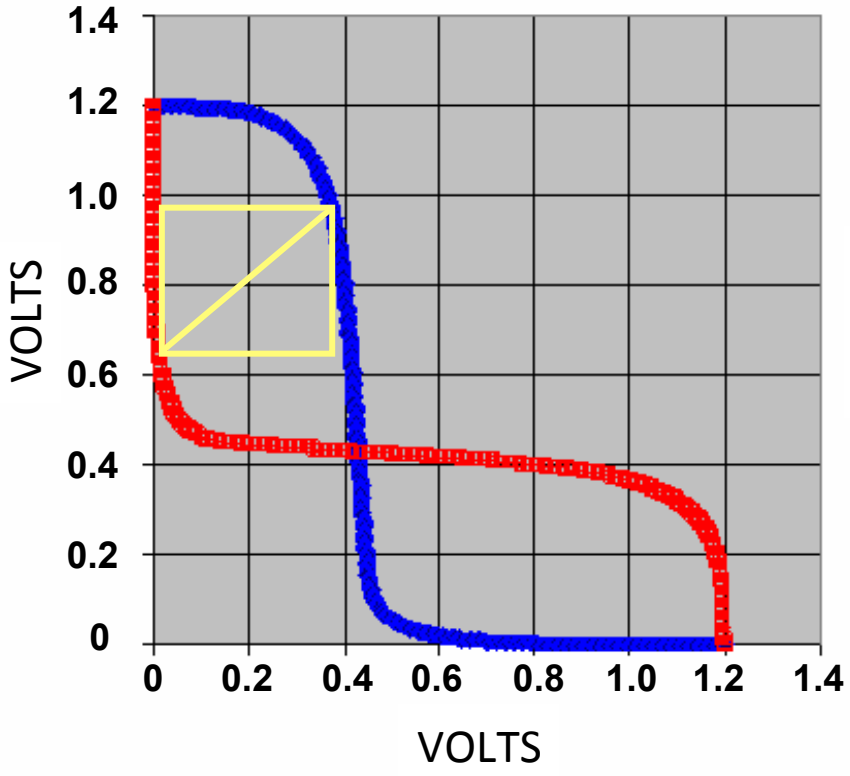
"Butterfly" Curve"



- SNM = maximum value of V_n that can be tolerated before cell changes state
- Additional margin is needed for stability over α-particle, crosstalk and PVT noise

SRAM Cell Stability Analysis

Retention Curve



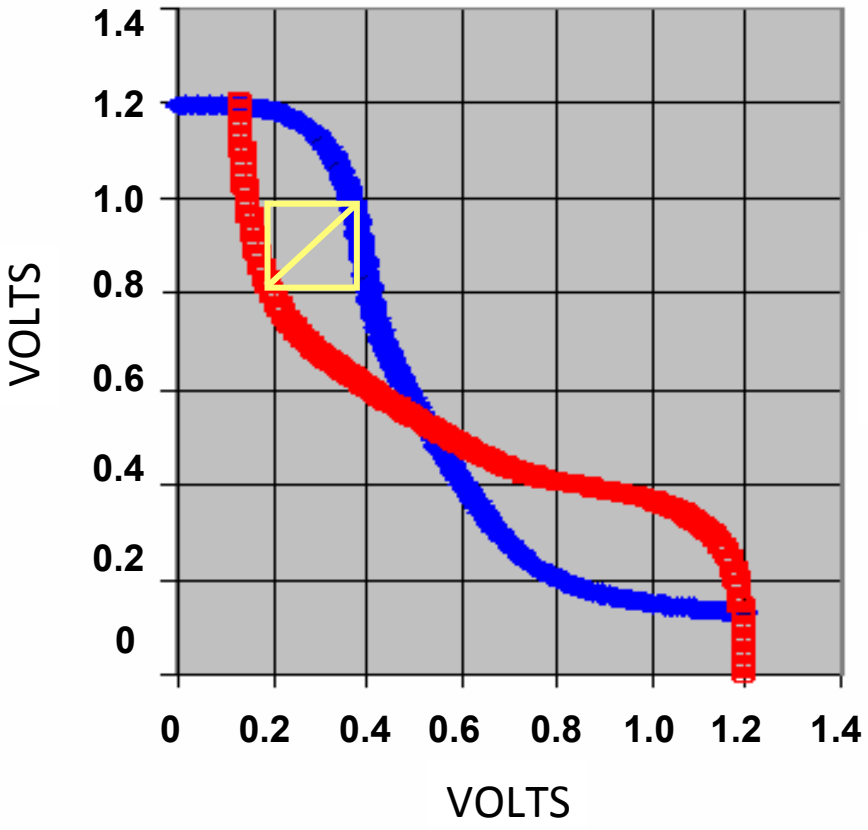
Large diagonal due to Pass Gate being "OFF"

Standby or "Data Retention" state when Wordline is not selected

→ Cell is stable

SRAM Cell Stability Analysis

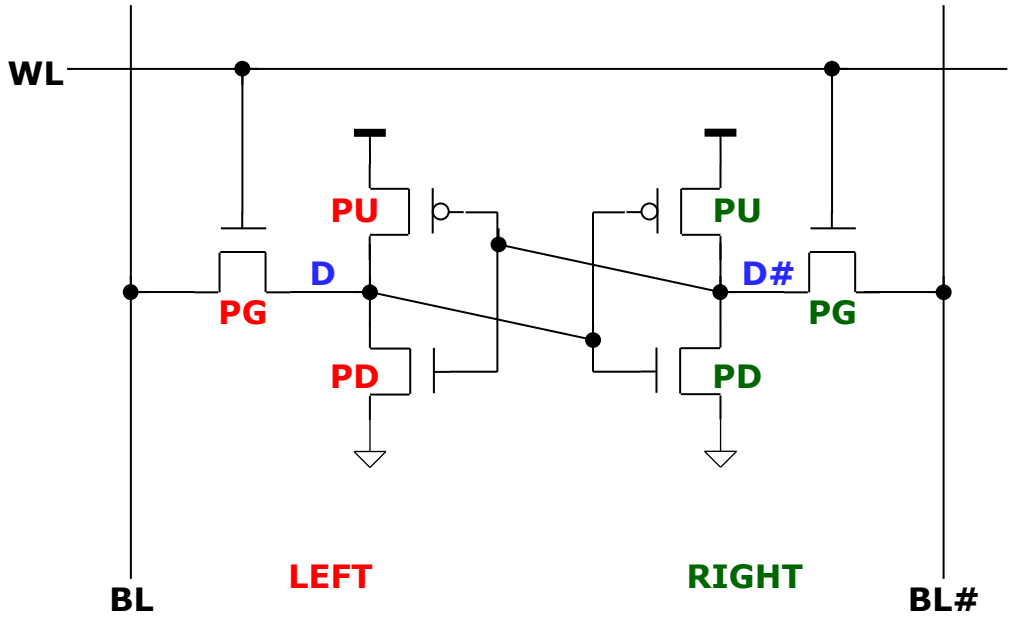
Stability Curve



Diagonal is reduced since Pass Gate is "ON"

The "maximum square" represents SNM of the bitcell

6T BIT-CELL DESIGN

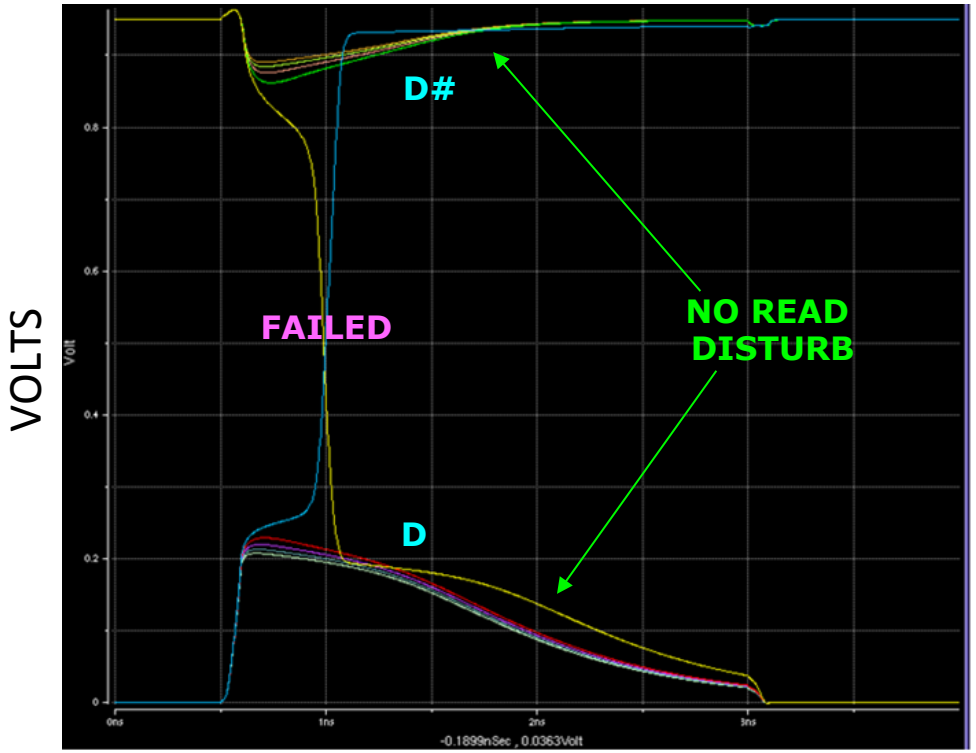


EXAMPLE: NORMALIZED DESIGN RATIOS

transistor	LEFT	RIGHT
Pass Gate (PG)	1.0	1.0
Pull Down (PD)	1.4	1.4
Pull Up (PU)	0.4	0.4

PD > PG for read stability
PG > PU to enable writes

READ DISTURB PROBLEM



- If the state nodes D (or D#) are 'forced' above (or below) the trip point of the **RIGHT** (or **LEFT**) inverter, the 6T cell state will be 'flipped' and thus creating a functional failure during a READ operation.
- What conditions are necessary to cause this? hint: look at the 6T cell design ratios.
- What can be done to mitigate a READ DISTURB condition?

READ and WRITE ASSIST

Without re-designing the 6T bit-cell, during a 'READ' operation:

- Lower the gate drive of the WORDLINE: effectively weakening the PG drive and making the PD > PG design requirement stronger; as a result, READ 'DISTURB' will be mitigated.

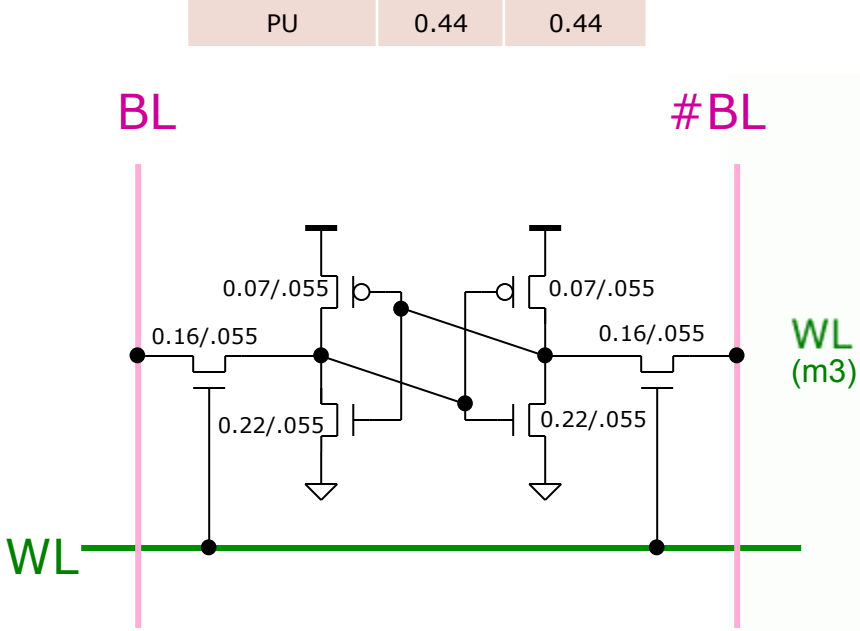
During a 'WRITE' operation:

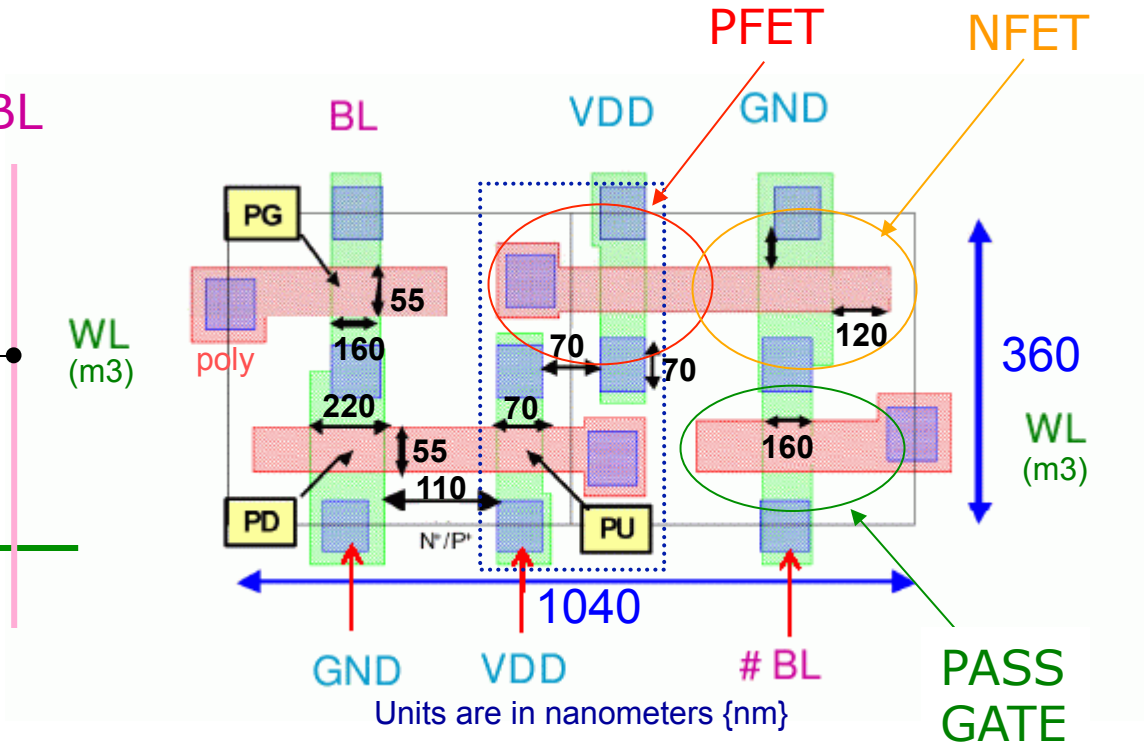
- Lower the array VDD power supply while keeping the WORDLINE driver's VDD supply high: effectively making the PG relatively stronger compared to the PU; this makes the design requirement of PG > PU stronger.
- Keep the array VDD supply steady, but 'Pulse' the gate of the PG transistor to >VDD.

Contemporary nm designs have made all 6 transistors equal in size (PG = PD = PU) and rely on READ and WRITE assist for proper operation. What is the benefit of making all 6 transistors the same size (and as small as possible)?

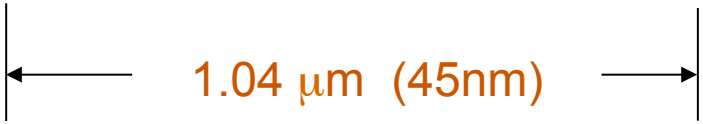
6-Transistor SRAM Cell Layout

transistor	LEFT	RIGHT
PG	1.00	1.00
PD	1.38	1.38
PU	0.44	0.44

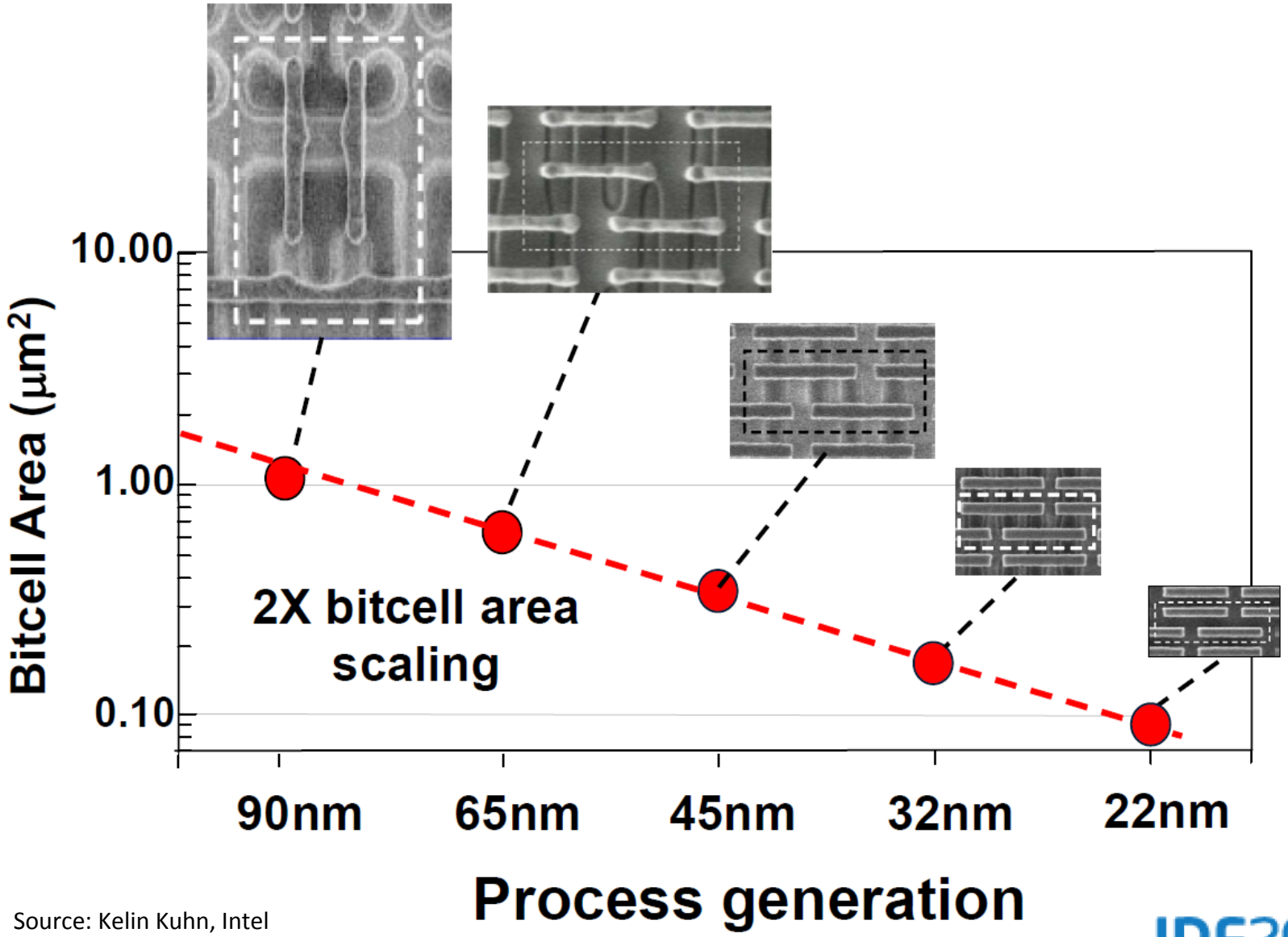




In 45nm CMOS, a typical
6T bit-cell area = 0.38 μm²



Moore's Law Scaling of the SRAM

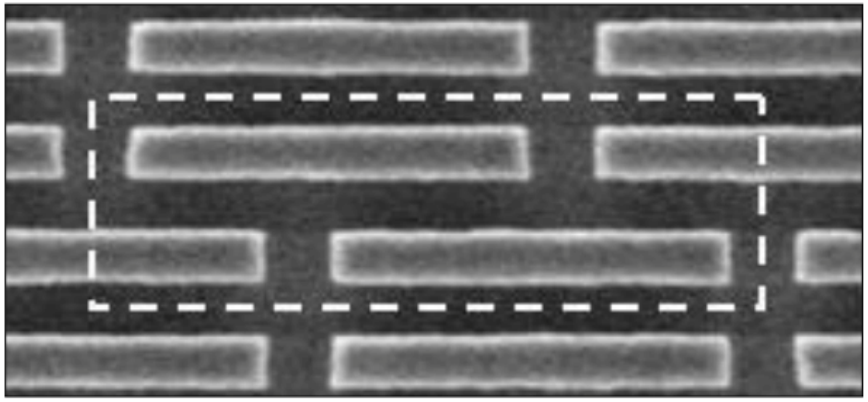


Source: Kelin Kuhn, Intel

IDF2011

SRAM Memory Cell Improvements [8]

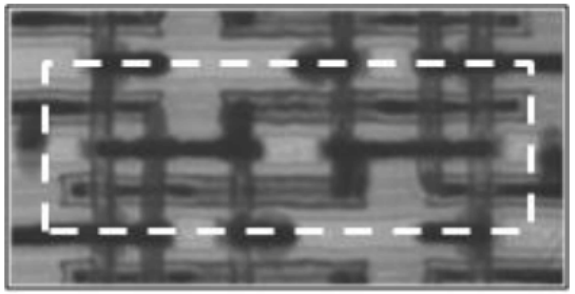
22 nm Process



.108 um^2

(Used on CPU products)

14 nm Process

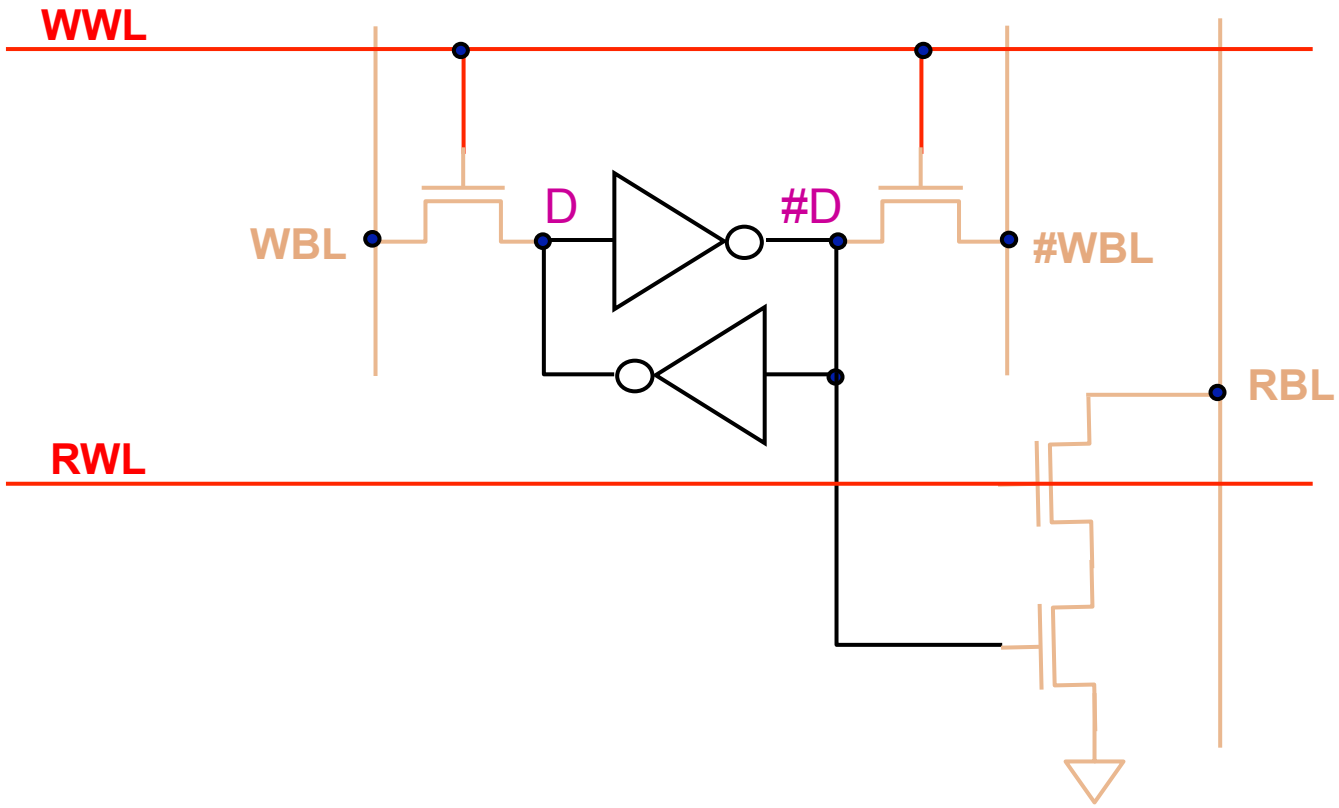


.0588 um^2

(0.54x area scaling)

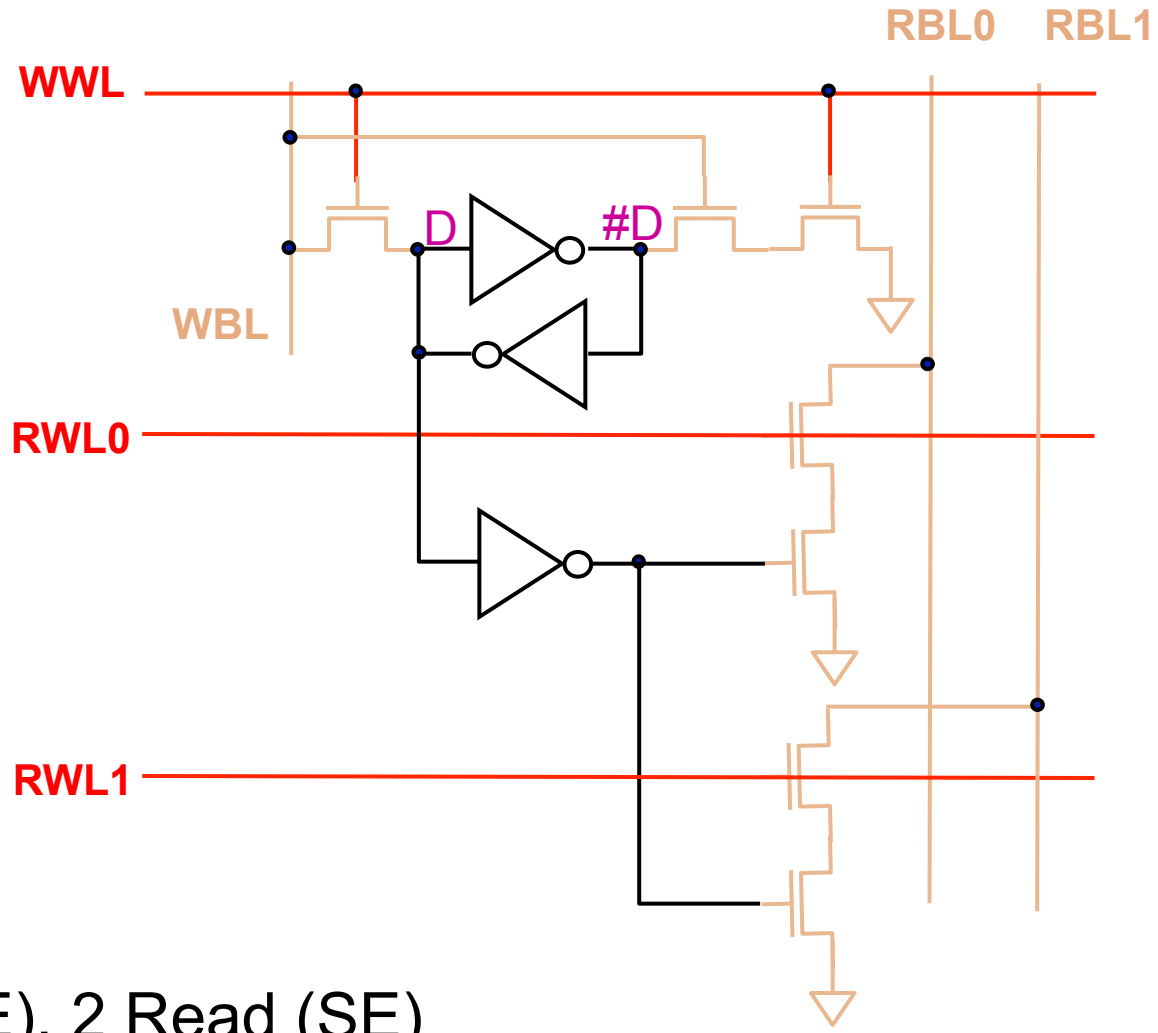
SINGLE-ENDED CELL

Separate word-lines and bit-lines for read and write ports.



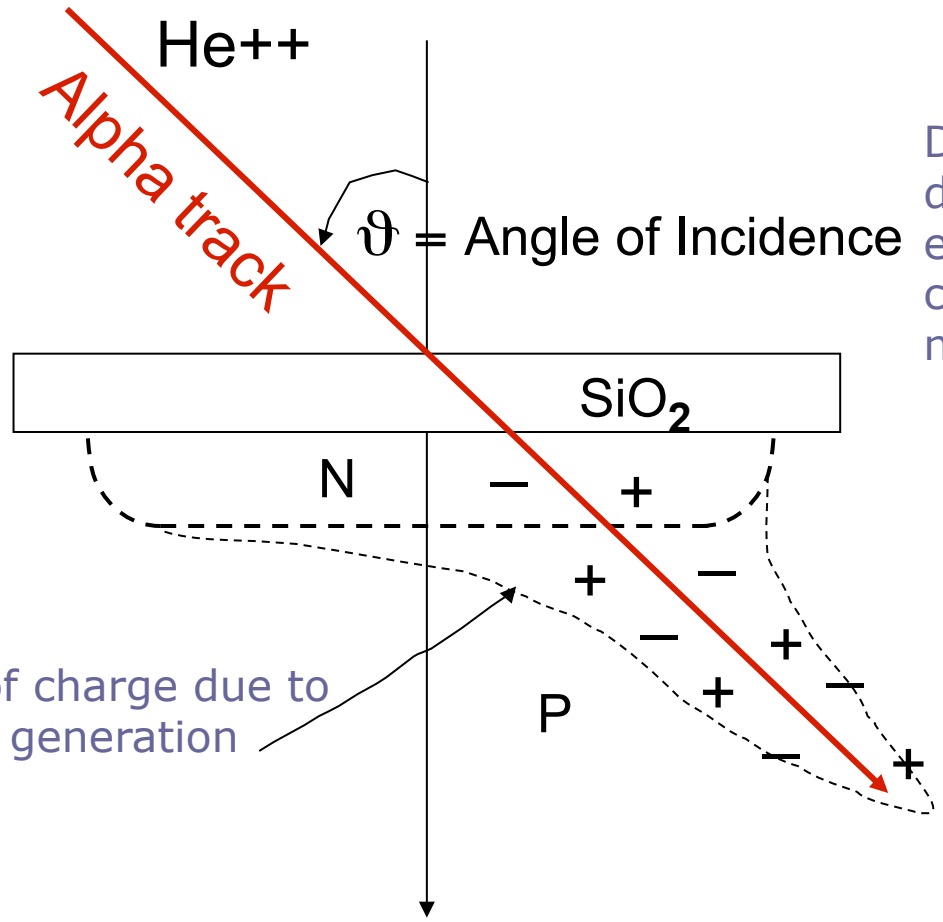
1 Write (DE), 1 Read (SE)

Multi-Port Memory Cell Types



1 Write (SE), 2 Read (SE)

ALPHA PARTICLES & FUNNEL EFFECT



Depletion region distortion results in enhanced charge collection by the hit node

Drift current causes the storage node to discharge

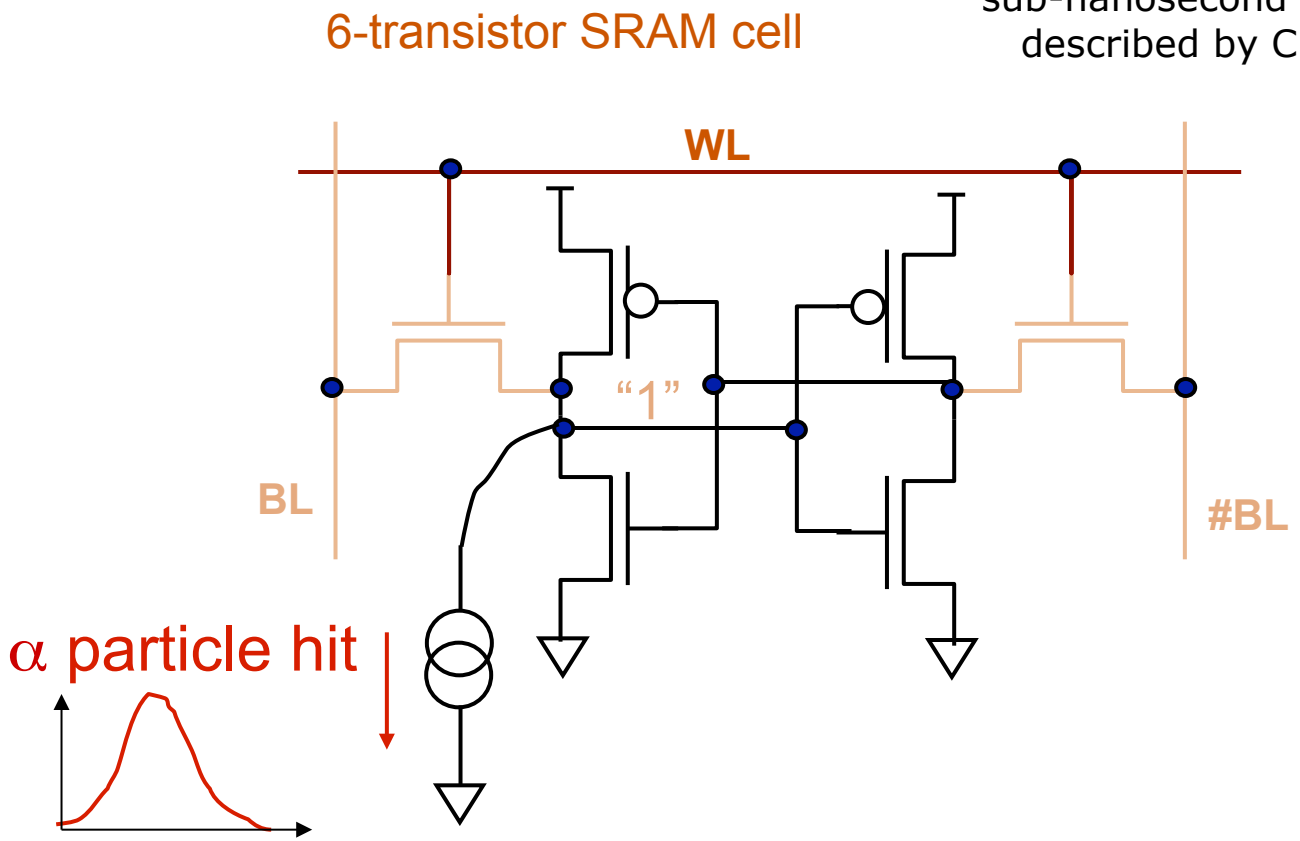
If the total charge exceeds Qcrit, then state of bitcell or latch will flip

Funnel of charge due to E-H pair generation

Qcrit = number of electrons which differentiates between a "1" and "0" (May & Woods) [1]

MODELING SOFT ERROR RATE

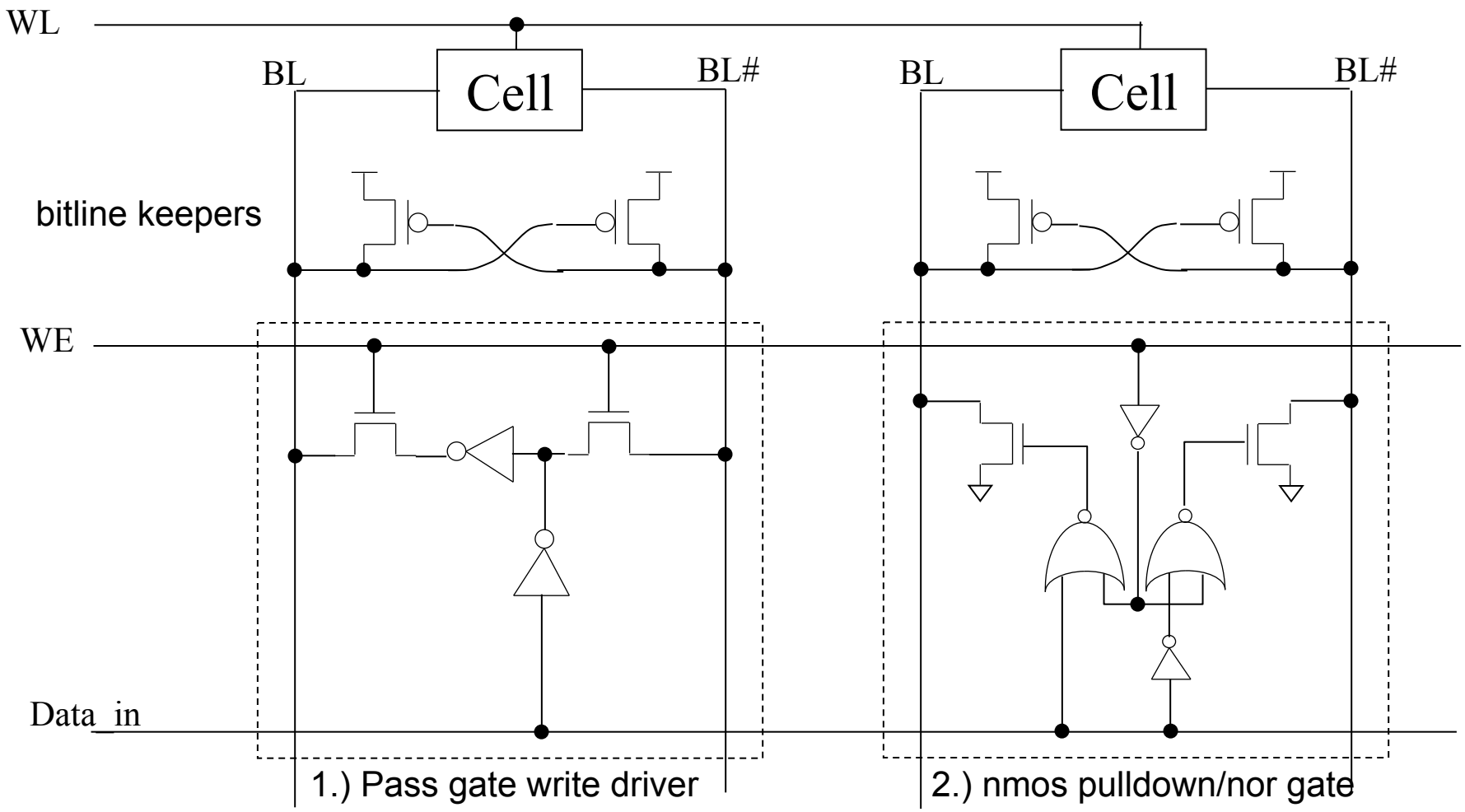
a-particle "hit" can be modeled as a sub-nanosecond current pulse as described by Chenming Hu [2]



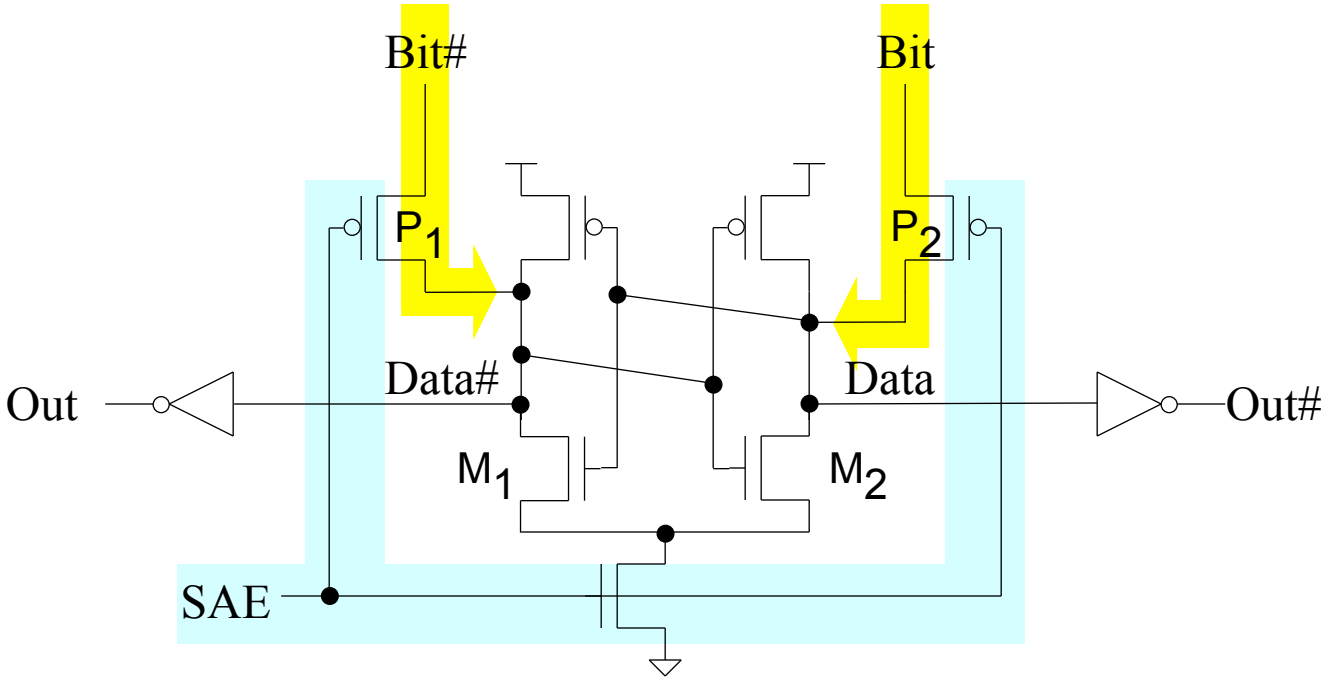
What can be done to this 6T cell to make it more robust to an a-particle hit?

WRITE CIRCUITS

2 example write circuit methods:



LATCHING SENSE AMPLIFIER

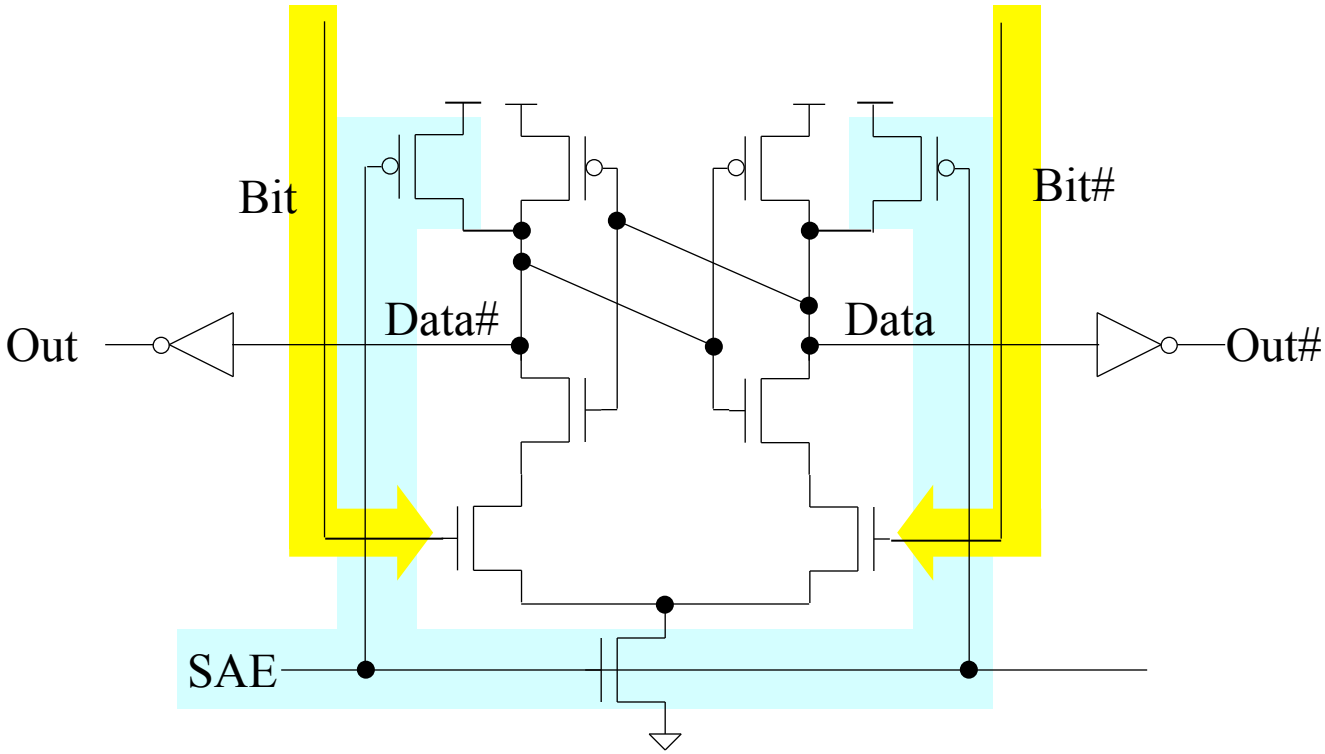


- If SAE is low; Data and Data# are connected to the bitlines.
- When SAE goes high; cross coupled inverters amplify and latch any voltage difference.

LATCHING SENSE AMPLIFIER

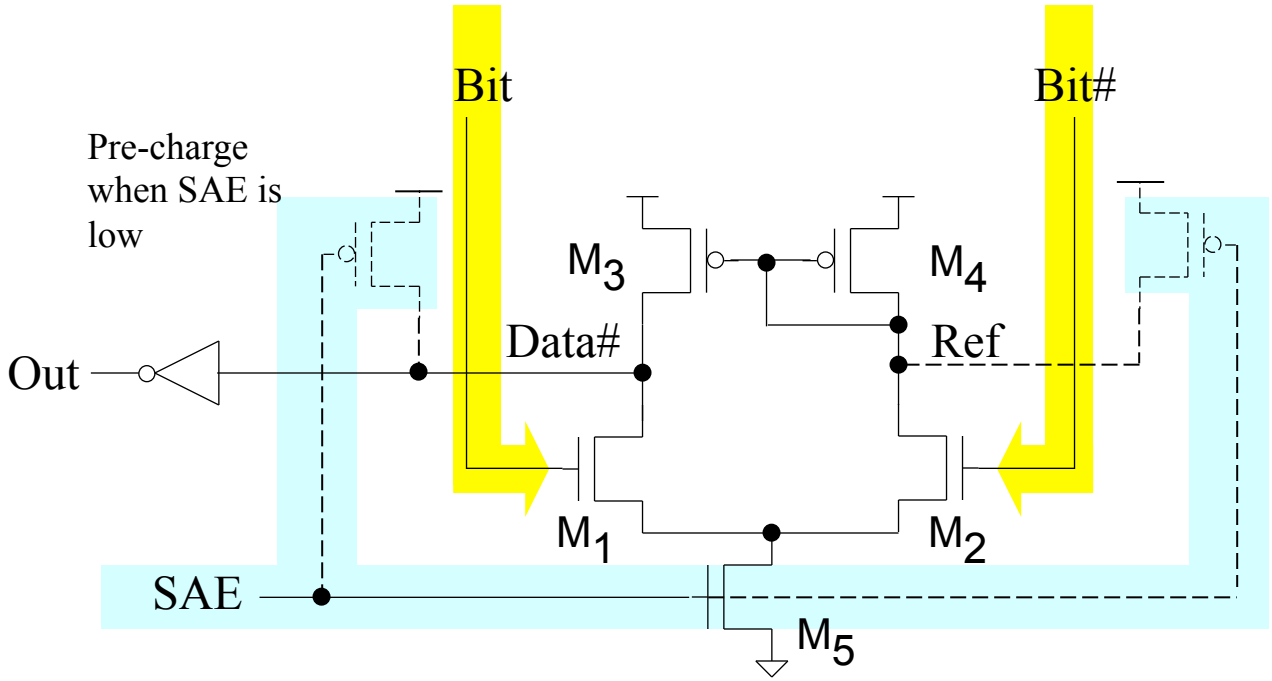
- Cross-coupled pair of inverters has high gain
  - Equalized to metastable region
- Internal nodes & bit-lines are pre-charged/equalized to VDD
- PMOS devices, P1, P2 are "on" allowing small differential voltage to propagate into sense-amp nodes; Data/Data#
- After a delay, sufficient bit-line differential is generated
- SAE is asserted and P1,P2 shut OFF
  - This decouples the highly capacitive load from the internal sense-amp nodes for FAST performance
  - M1, M2 turn "on", but one side is slightly stronger than the other
- This leads to slight instability, but positive feedback re-enforces the proper state
- Provide differential outputs & latching function (as long as SAE is high)

DE-COUPLED SENSE AMPLIFIER



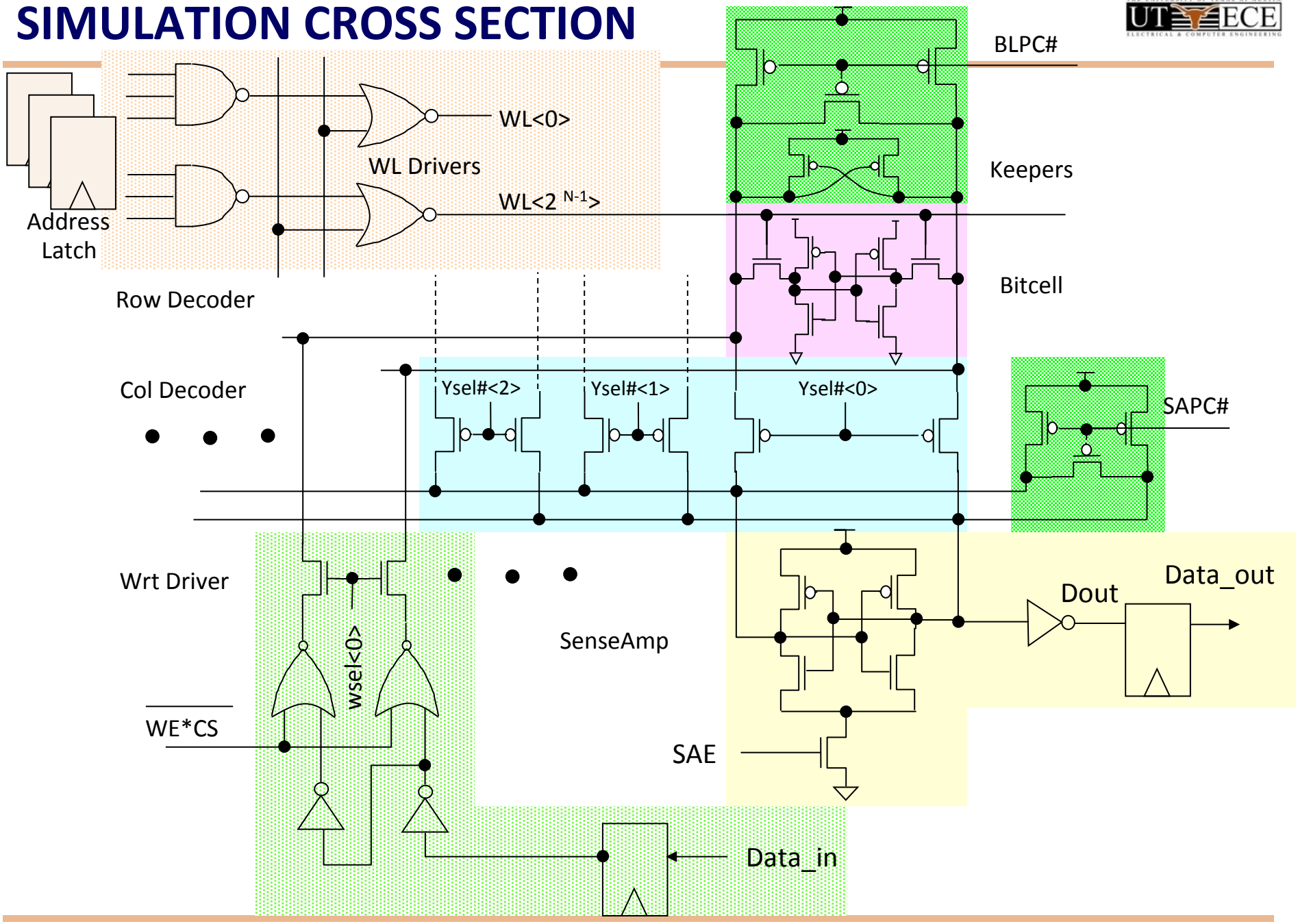
- With SAE low; Data and Data# are pre-charge high
- When SAE goes high; source-coupled pair acts as differential amplifier
- Cross coupled inverters amplify and latch any voltage difference

CURRENT MIRROR SENSE AMPLIFIER

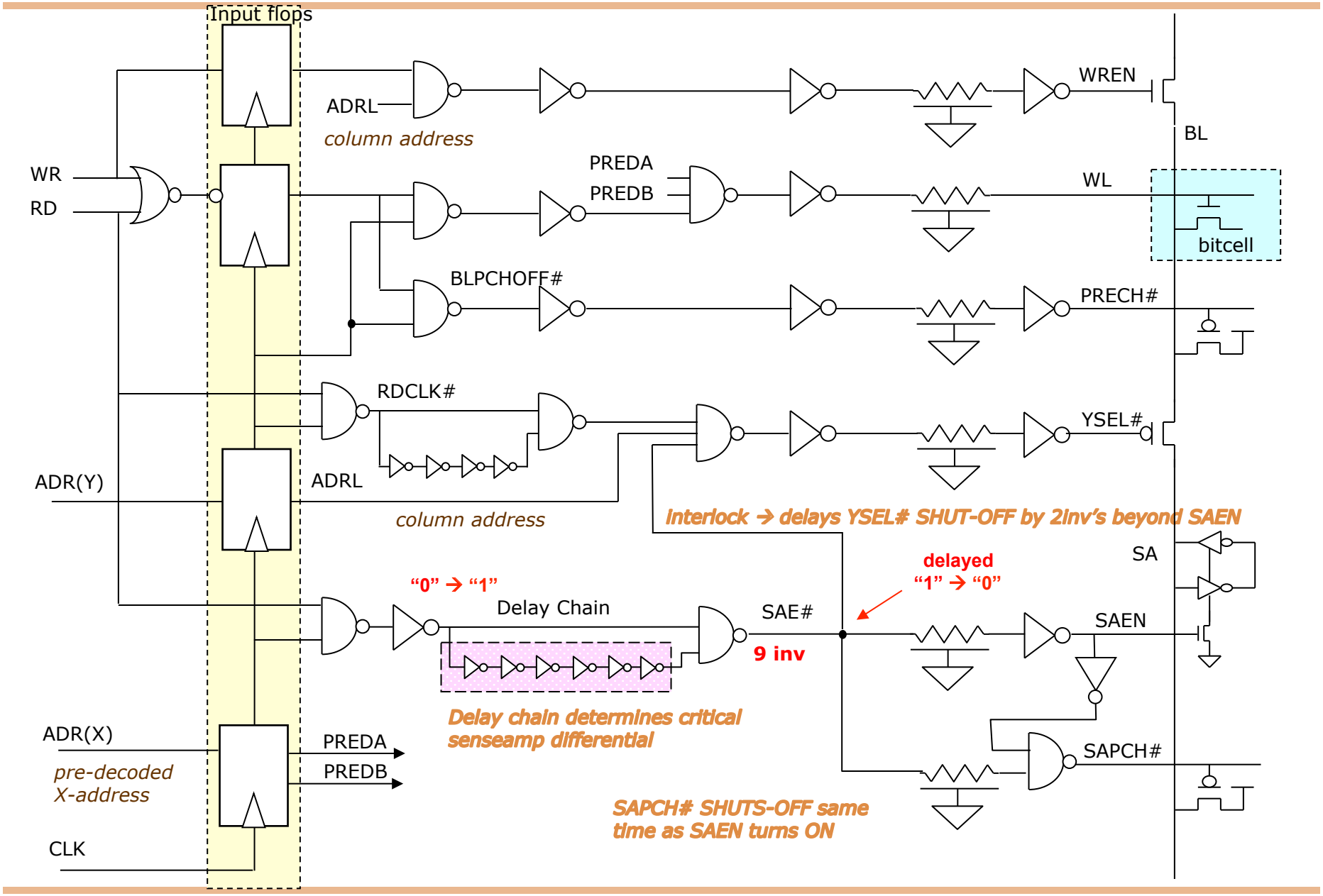


- M1 and M2 form a source-coupled pair
- M2 current is "mirrored" by (M3,M4) and compared to M1's drain current
- Voltage difference on bit-lines causes differential pair to allow current steering to one side
- Current difference is translated into high voltage gain; large swing at node Data#
- More detailed description in Weste & Harris "CMOS VLSI Design"

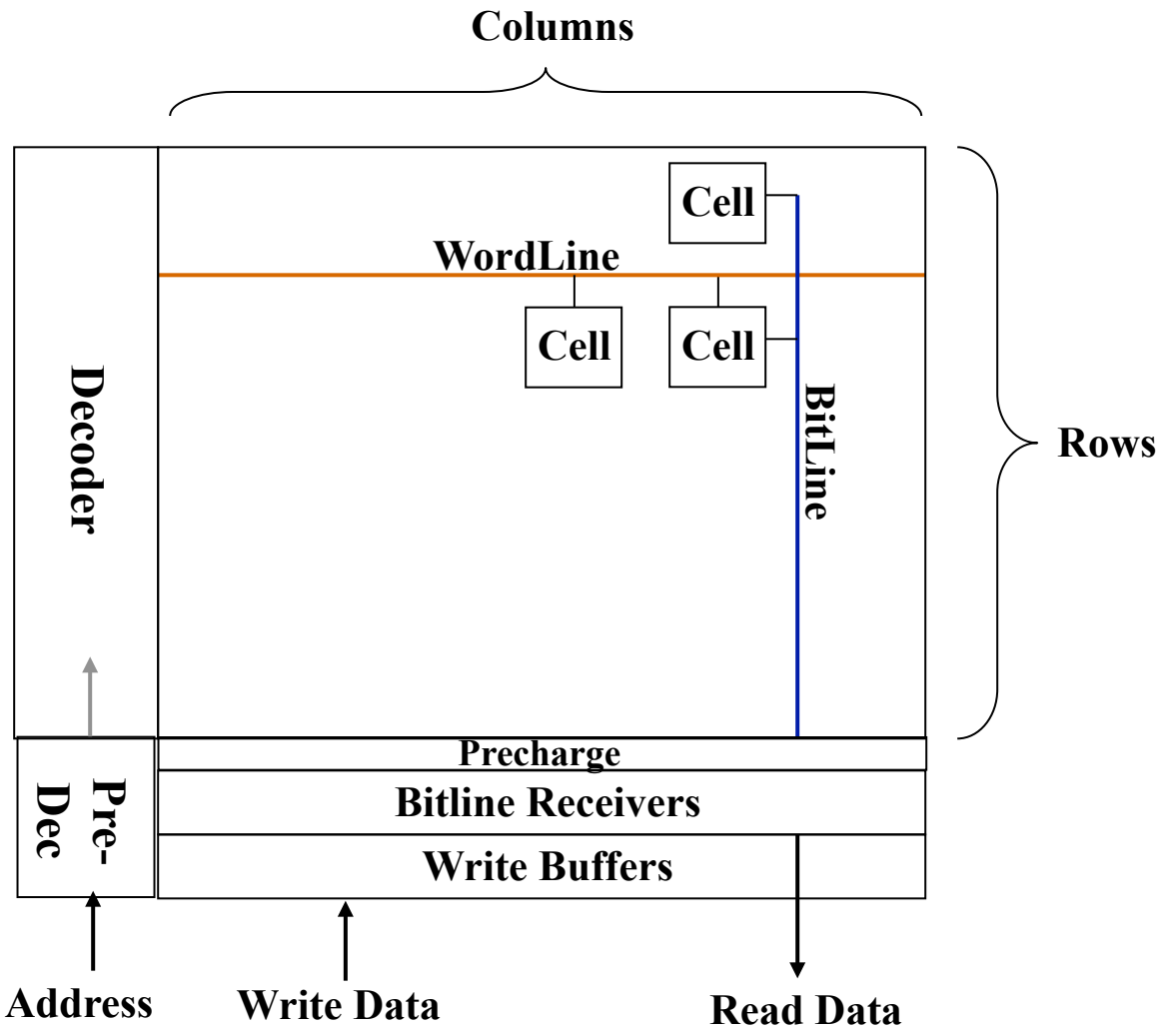
SIMULATION CROSS SECTION



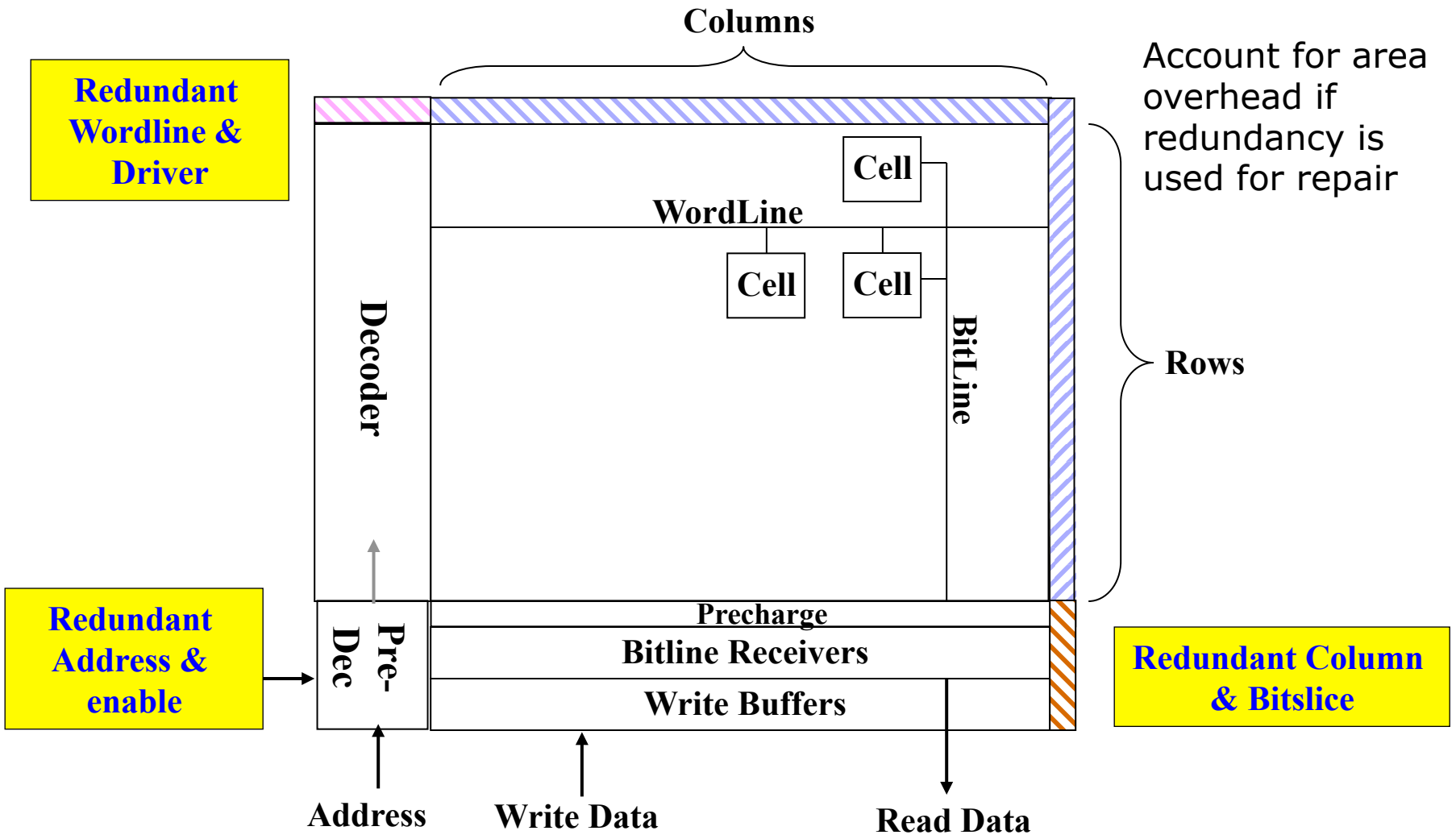
SELF-TIMER CIRCUIT



BASIC ARRAY LAYOUT



ARRAY REDUNDANT ELEMENTS



Sample SRAM Layout Using a Memory Compiler

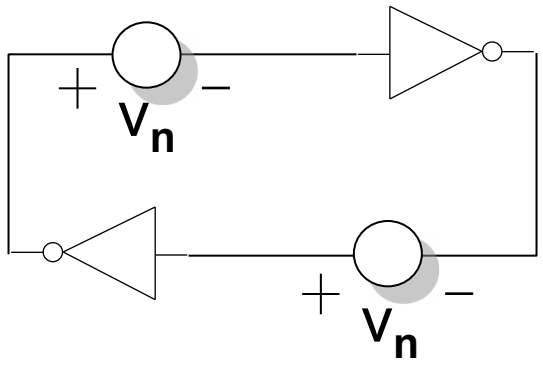
REFERENCES

1. T.C. May and M.H. Woods: Alpha-Particle-Induced Soft Errors in Dynamic Memories, IEEE Trans. Electron Devices, VOL. ED-26, NO.1, January 1979
2. C. Hu : Alpha-Particle-Induced Field and Enhanced Collection of Carriers, IEEE Electron Device Letters, VOL. EDL-3, NO.2, February 1982
3. E. Seevinck *et al.:* Static-Noise Margin Analysis of MOS SRAM Cells, IEEE Journal of Solid State Circuits, VOL 22, NO. 5, November 1987
4. Gian Gerosa *et al*.: 250MHz 5-W POWERPC MICROPROCESSOR, IEEE Journal of Solid State Circuits, VOL 32, NO. 11, November 1997
5. C. Nicoletta *et al.:* A 450-MHz RISC Microprocessor with Enhanced Instruction Set and Copper Interconnect, IEEE Journal of Solid State Circuits, VOL 34, NO.11, November 1999
6. http://www.eecg.toronto.edu/~pagiamt/cam/camintro.html
7. http://www.eecg.toronto.edu/~pagiamt/cam/references.html
8. Mark Bohr, Intel Senior Fellow, Intel's 14nm Technology Announcement, August 11, 2014

BACKUP

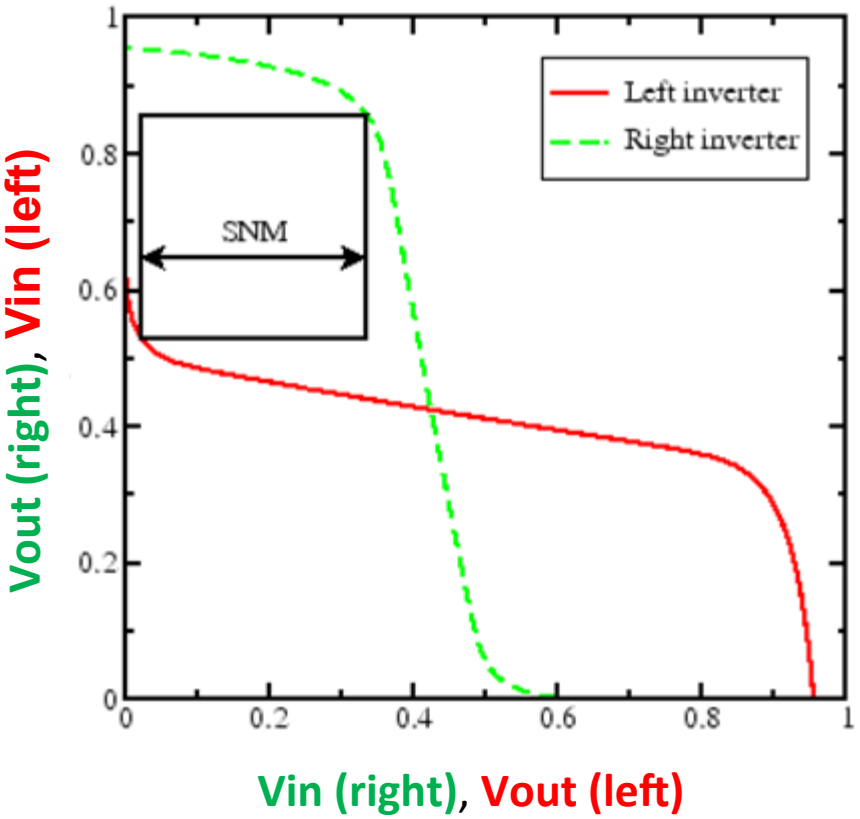
SRAM Cell Stability Analysis

- Bitlines are precharged to VCC → this is the critical situation because the nmos access device "shunts" the pmos load device; thereby reducing the gain of the inverters
- Static noise voltage sources are inserted into cross-couple path between inverters



- SNM is defined by the maximum value of Vn that can be tolerated before changing state; sweep noise voltage from 0V to the point where differential collapses to zero
- Include temperature, voltage and process variation using a Monte Carlo simulator

STATIC NOISE MARGIN



SOFT ERROR RATE BACKGROUND

- There are 2 categories of system failure:
  - hard failure (permanent failures that require replacement)
  - soft failure (non-permanent random system failures)
- Cause of failures could be noise, power glitches, design margins, etc
- In large memory systems, soft errors are mostly due to radiation
- In 1978, May & Woods[5] (Intel) found radioactive materials in memory packages emitting alpha particles which can generate sufficient charge to switch the state of stored charge in DRAMs
- Minute traces of radioactive elements can be found in alumina-based ceramics, zirconia & silica fillers used in packaging
- Another potential source of alpha particles is from cosmic radiation
  - High energy particles from cosmic rays can have energies greater than 1GeV
  - Alpha particle energies typically range from 0.1 to 10 MeV

ALPHA PARTICLES

- An alpha-particle is a doubly charged helium nucleus (2 protons, 2 neutrons) that is generated during radioactive decay of high-Z atoms
- More than 300 known alpha-emitting nuclides:
  - Uranium(238), Thorium(232) can be found in package materials for semiconductors
  - Radioactive decay of $U^{238} \rightarrow Th^{234} + He^{4}$ until decays to stable Pb^{206} (8 alphas are generated)
  - Thorium generates 6 alphas as it decays from Th^{232} to stable Pb^{208}
- Alpha particles interact with silicon to generate an ionization trail of electron-hole pairs
- The amount of electron-hole pairs generated depends on the particle's initial energy (~3.6eV per e-h pair)