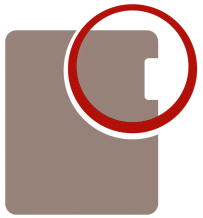
National Taiwan University

Memory Testing

Jiun-Lang Huang
ICDA/GIEE, NTU

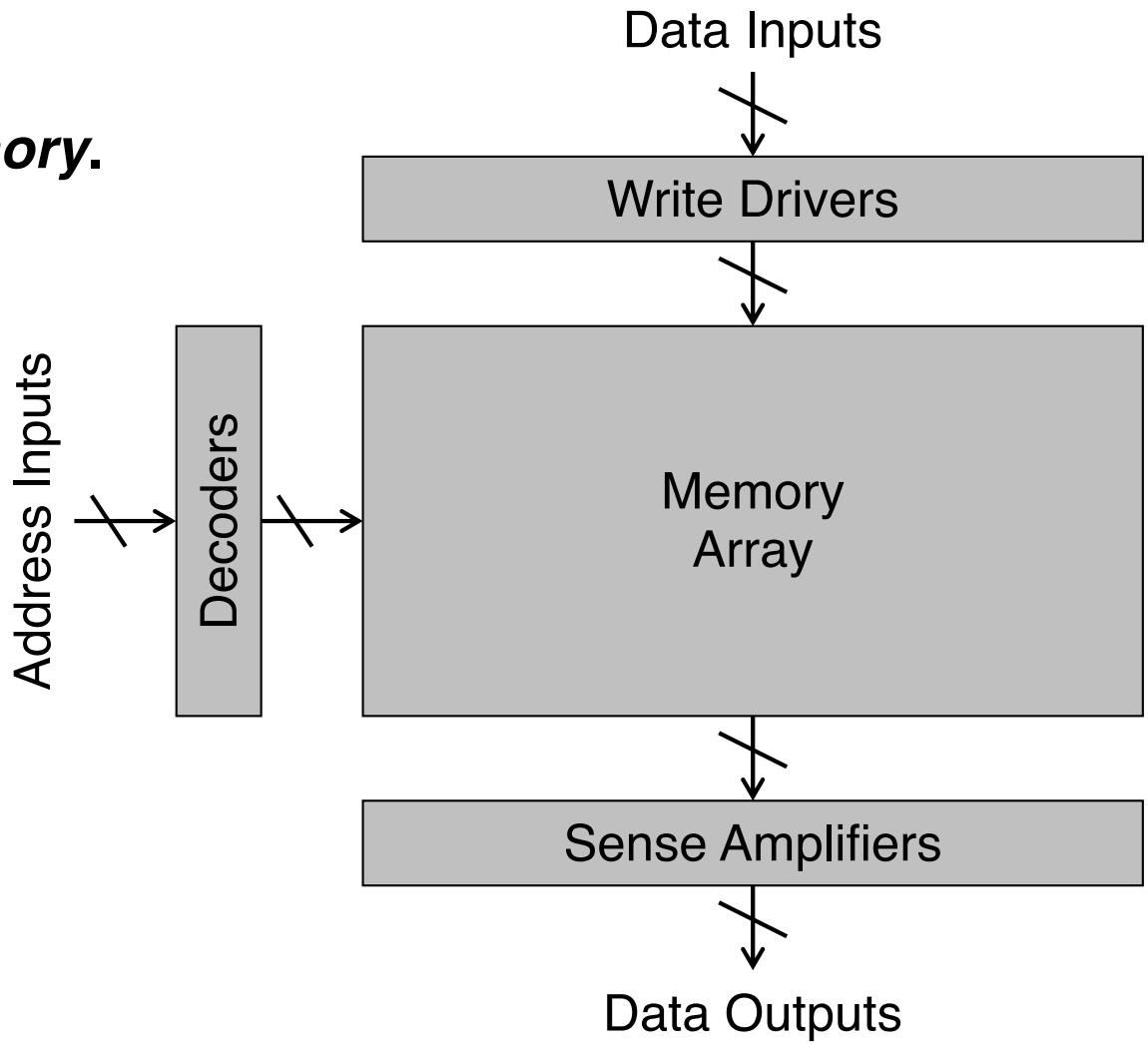
Outline



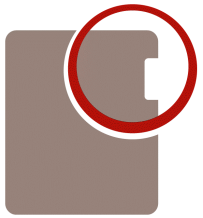
- **Overview**
- Design and test considerations
- Memory testing
- Memory self test

A Simple Memory Block Diagram

Our focus is:
***embedded memory*.**

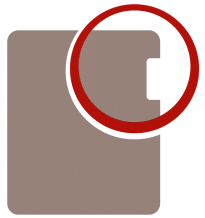


Memory Testing Complexities



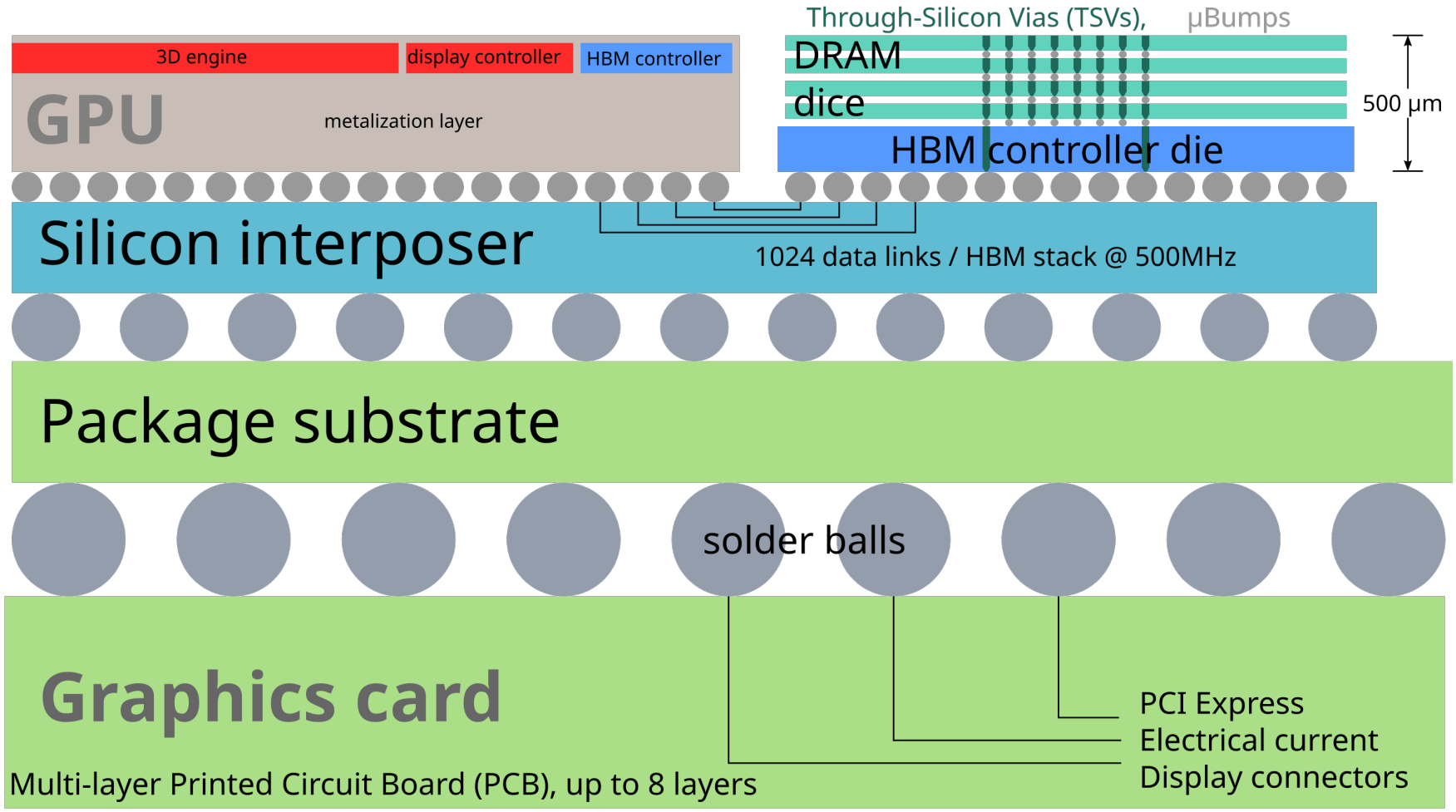
- Increasing density (the Moore's law).
  - More to test, more test time.
- Addition of logic to memories.
  - Random logic or memory testing techniques alone do not achieve high fault coverage for both.
- Memories are more deeply embedded.
  - 40 different memory designs on a single chip!!
  - Direct access is simply impossible.

Memory Testing Complexities - cont'd



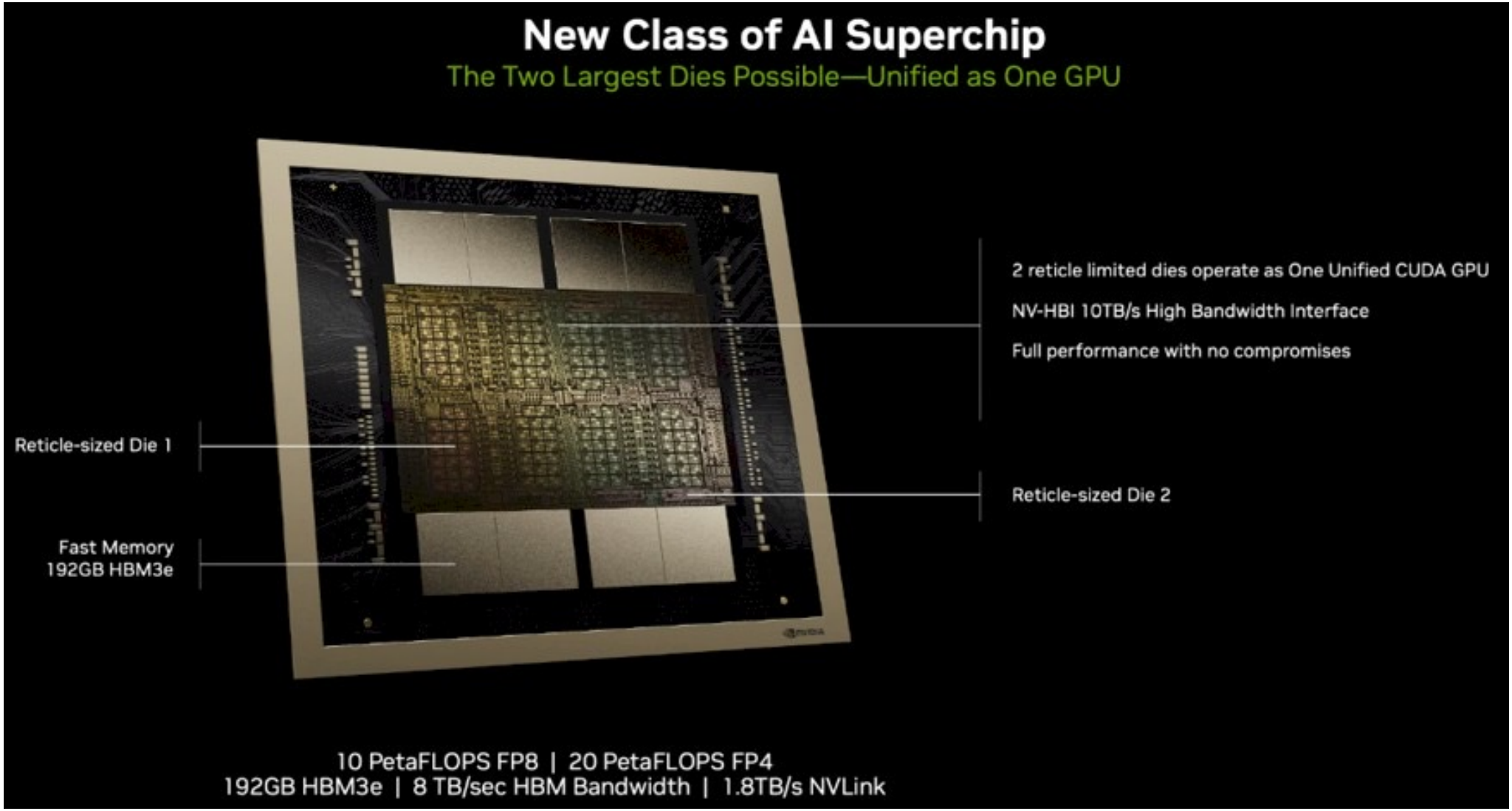
- Increasing number of memory types.
  - Volatile: DRAM, SRAM, CAM, TCAM, …
    - Multi-port, pseudo multi-port.
  - Non-volatile: Flash, EEPROM, FeRAM, MRAM, OUM.
- Different process technologies.
  - CMOS, SOI, SiGe, …
  - Strained silicon, silicon on nothing, …
- Redundancy
  - To enhance yield for large memories.
  - Involves locating the faults and allocating the redundant elements.

High Bandwidth Memory (HBM)



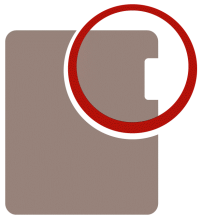
https://commons.wikimedia.org/wiki/File:High_Bandwidth_Memory_schematic.svg

NVIDIA Blackwell



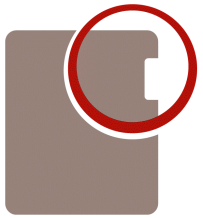
http://www.nextplatform.com/wp-content/uploads/2024/03/nvidia-blackwell-specs.jpg

To BIST or Not?



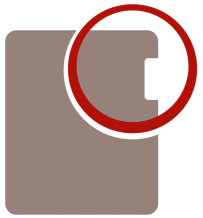
- For embedded memories, BIST is the only practical solution.
  - Difficulties in accessing deeply embedded memories.
  - Need of at-speed testing.
    - Reducing test time.
    - Detecting subtle errors.
    - Better quality test.
    - Practically impossible w/ external testers.
  - Pushes the test into the design phase.
  - Better test re-use at each step in the manufacturing and use process.

Importance of Test Quality



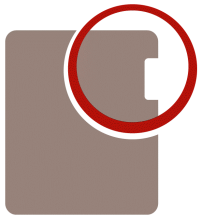
- A defective memory can pass a poor quality test.
  - A memory that passes test is only as good as the test applied!
- Ignorance, i.e., testing without knowledge, may result in an artificially high yield but the result will be disastrous.

Outline



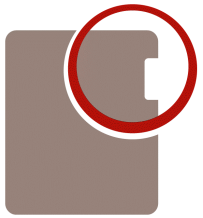
- Overview
- Design and test considerations
  - **SRAM**
  - Multi-port memories
  - DRAM
- Memory testing
- Memory self test

SRAM



- The mainstream of embedded memories.
  - Highest speed.
  - Can be designed for low power application requirements.
  - The easiest to use.

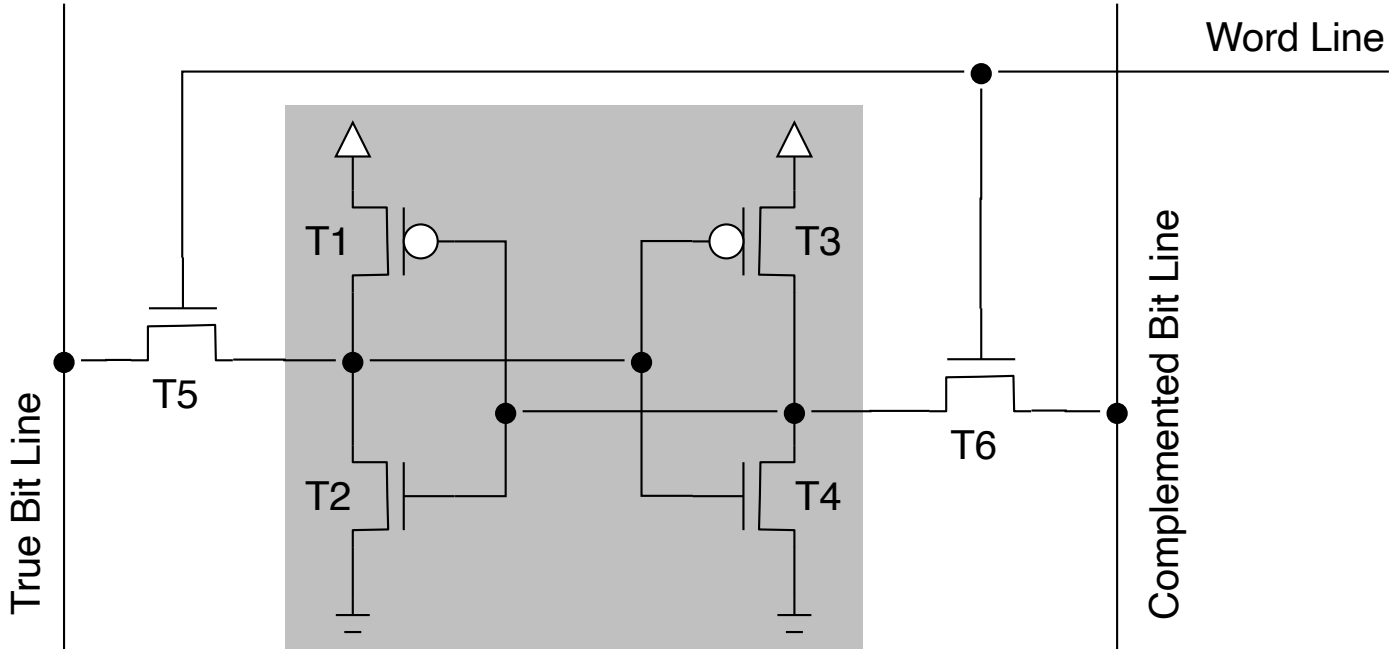
SRAM



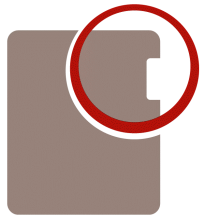
- **SRAM cell design**
- Read data path
- Write driver circuit
- Decoder circuitry
- Layout considerations
- Redundancy

A 6-T SRAM Cell

- T1 and T3: pull-up devices.
  - Replaced by resistors in 4-T cells.
- T2 and T4: pull-down devices.
- T5 and T6: transfer devices.

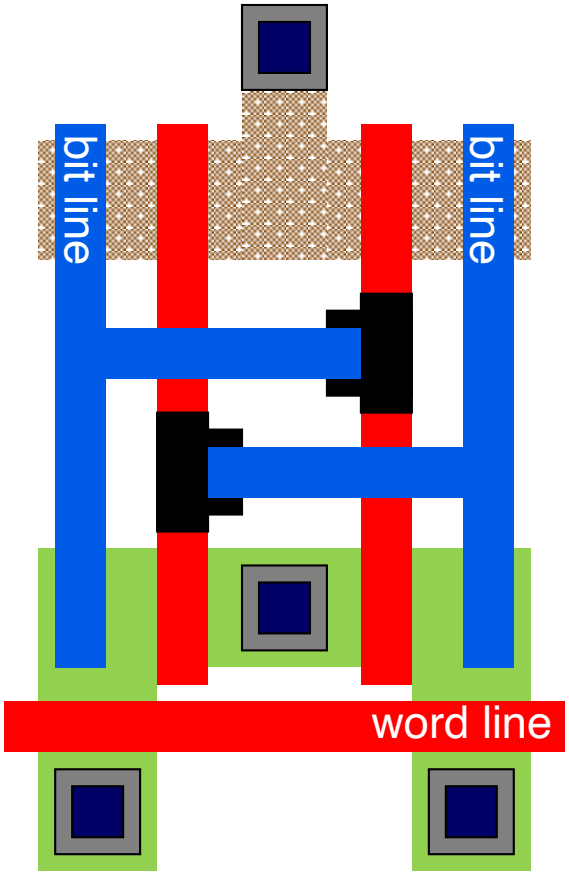


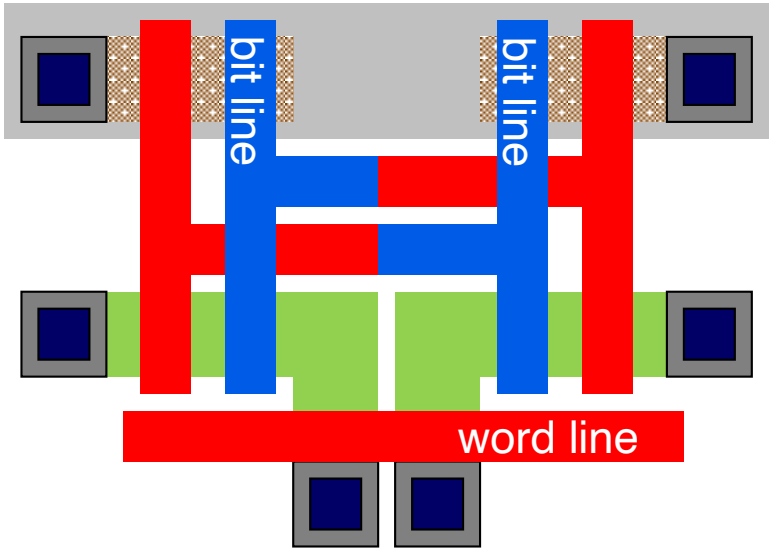
Cell Stability - the Beta Ratio



- The ratio of the strength of the pull-down transistor divided by the strength of the transfer transistor.
  - Aka the “beta ratio.”
  - $\beta = (W_{eff\_pd}/L_{eff\_pd})/(W_{eff\_tfr}/L_{eff\_tfr})$
- A beta ratio of 1.5 to 2.0 is typical.
  - A beta ratio below 1.0 indicates that each time the cell is read, it is disturbed as well.
  - A defect-free SRAM cell must have a non-destructive read operation.

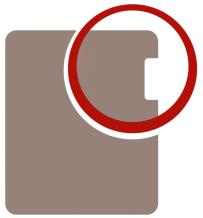
6-T Cell Layout





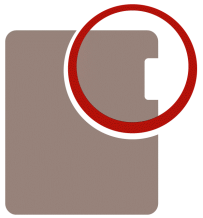
- Area is the major concern.
- The same defect mechanism causes different fault effects.

Remarks



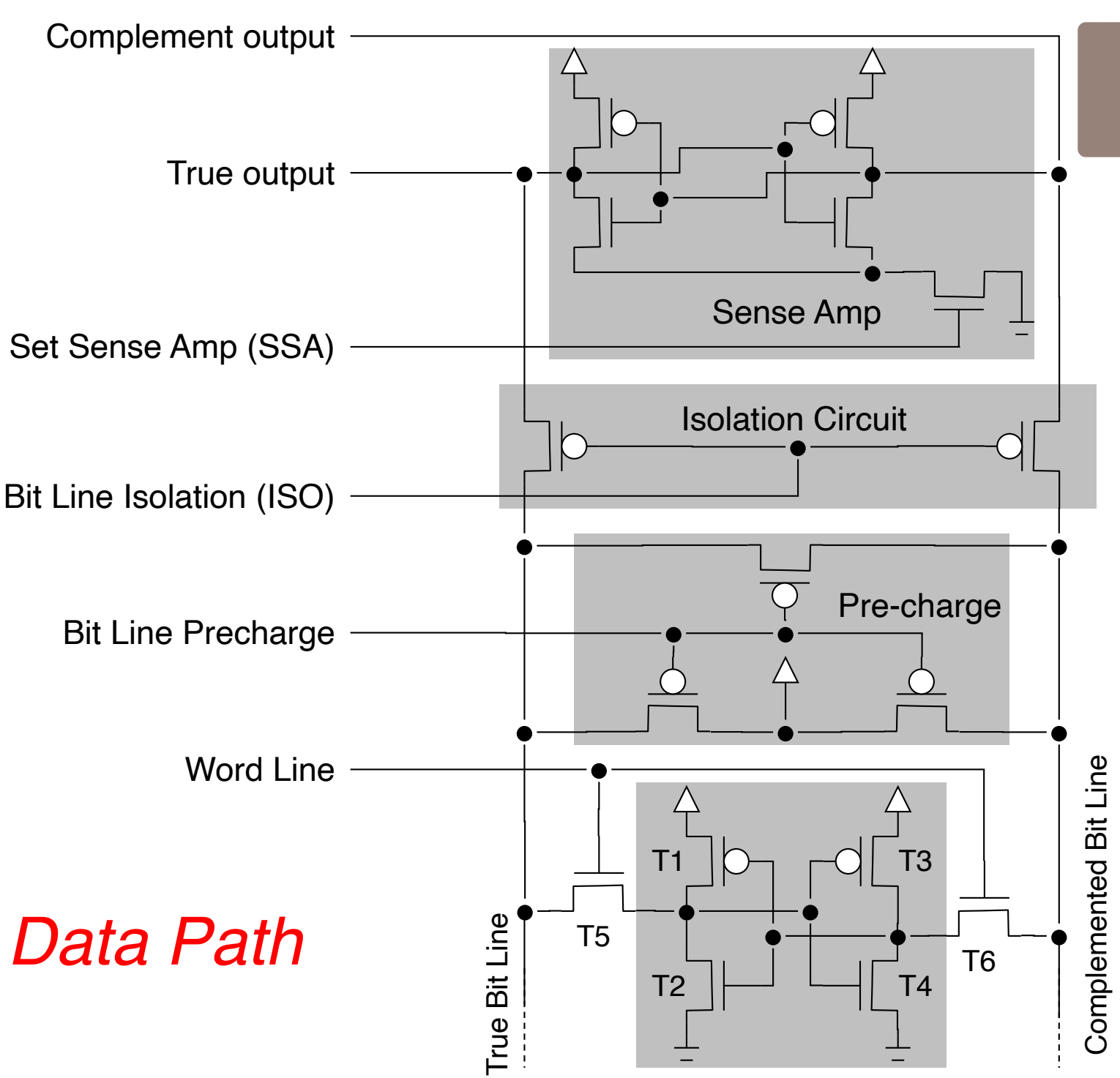
- The same defect mechanism may cause different faults.
  - A resistive ground causes imbalance in the single ground contact layout.
  - A defective shared V_{dd} contact will fail a group of cells.
- Different fault models and test patterns should be used for different layout designs.

SRAM

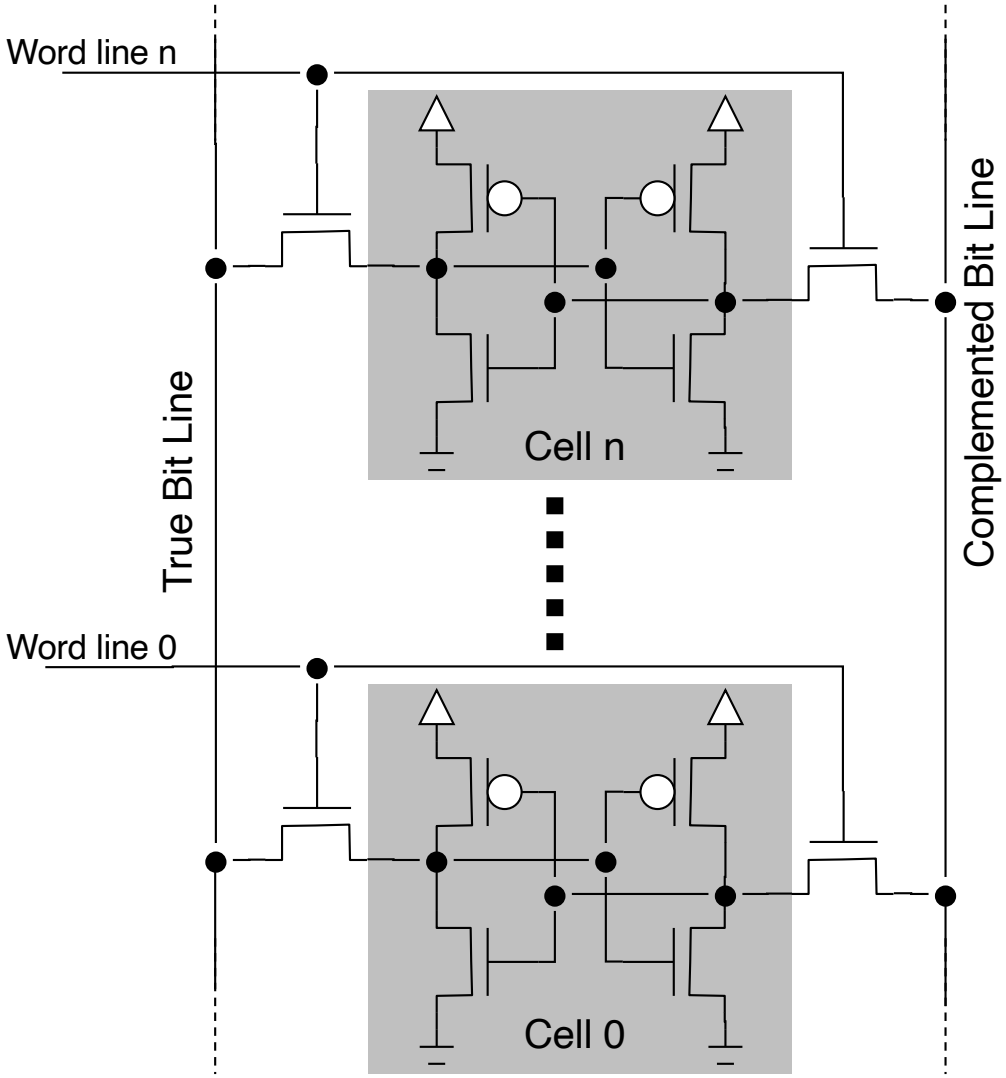


- SRAM cell design
- **Read data path**
- Write driver circuit
- Decoder circuitry
- Layout considerations
- Redundancy

Read Data Path

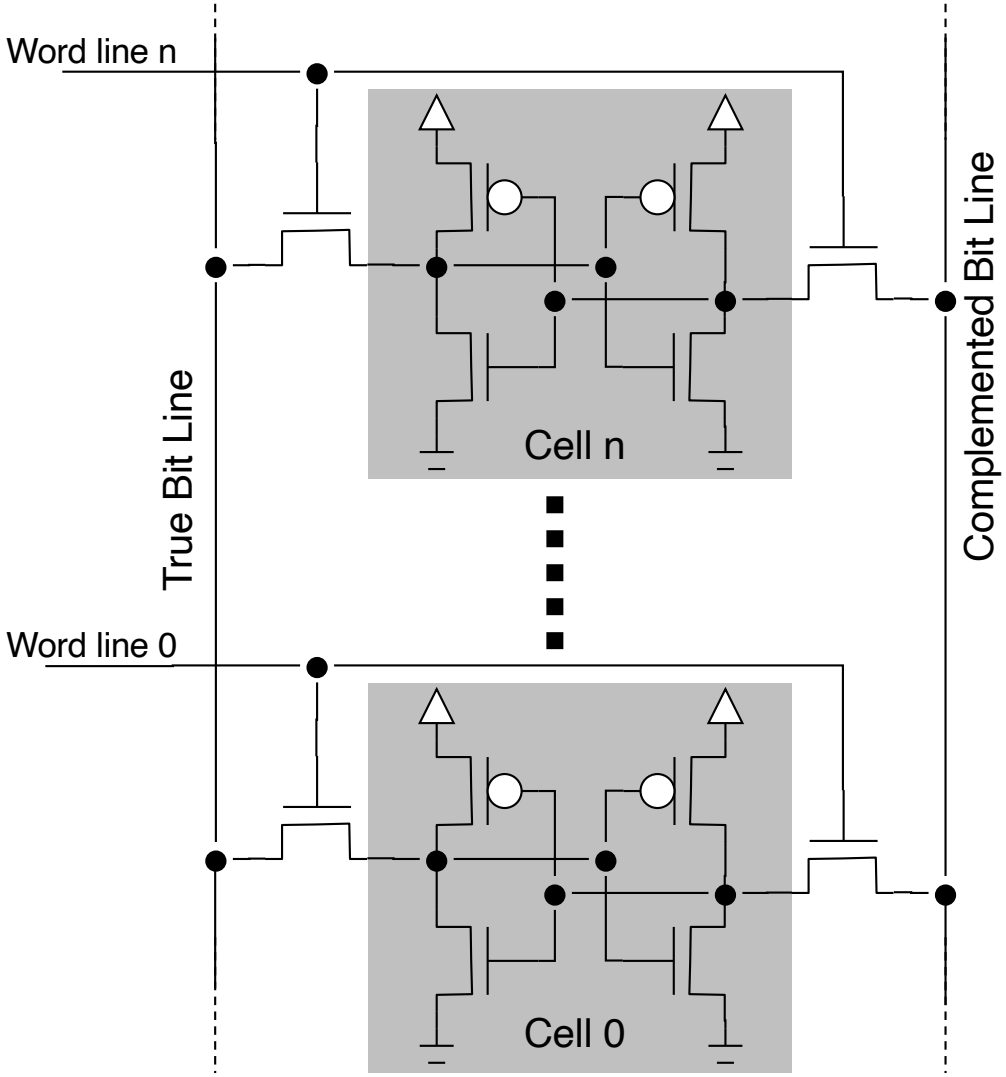


Bit Lines



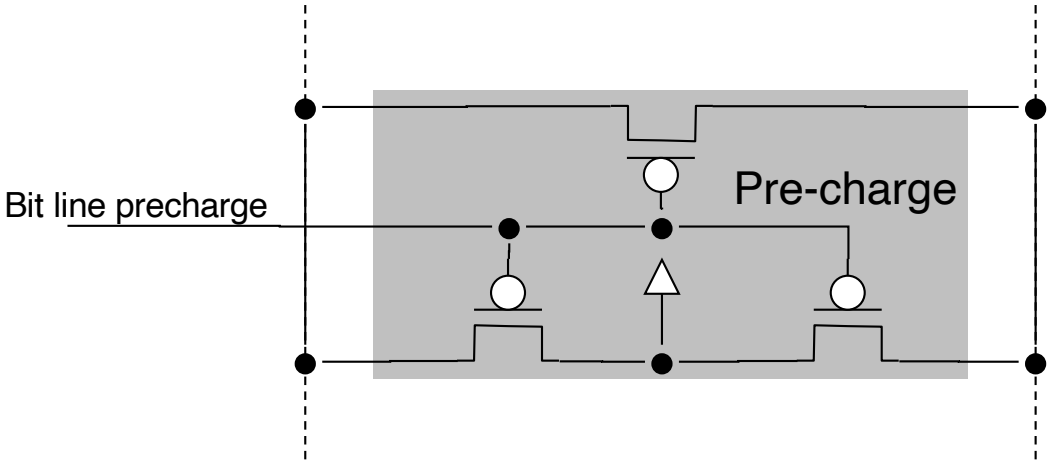
- Cells are joined along a column by bit lines.
- High speed SRAMs always use a pair of bit lines for reading and writing.
  - On a read operation, the bit lines only swing a small amount.
- *n* is normally large, often being 256, 512, or 1024 in value.

Bit Lines - Read Operation



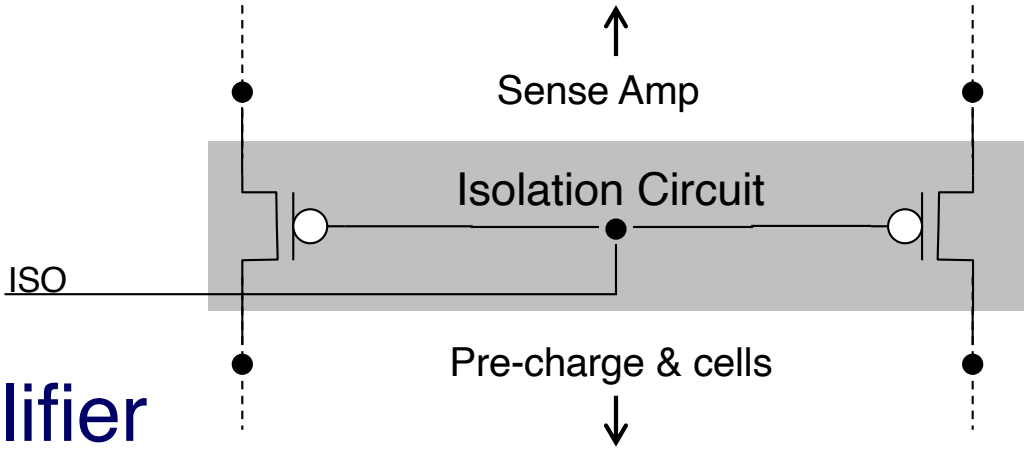
- Normally, the bit lines are pre-charged to Vdd.
- The word line corresponding to the address will go high and turn on the transfer devices.
- One of the bit lines will be pulled down.
  - The other remains unchanged.
  - The difference between the pair may only be 100 mV.

Pre-Charge



- Pre-charging to Vdd is easier and the norm in SRAM designs today.
- During a read operation, the pre-charge circuit is normally turned off for a column that's being read.
- The pre-charge is often left on for the columns that are not being read.
  - Half-select state: pre-charge on and word line high.

Isolation Circuitry



- During a read operation, the bit lines are isolated from the the sense amplifier once sufficient signal is developed.
- The purpose is to isolate the sense amplifier from the bit line's load and thus speed up the sense amplifier circuit operation.
- A sense amp may be shared by several columns. In this case, the isolation circuit is replicated to form a multiplexer.

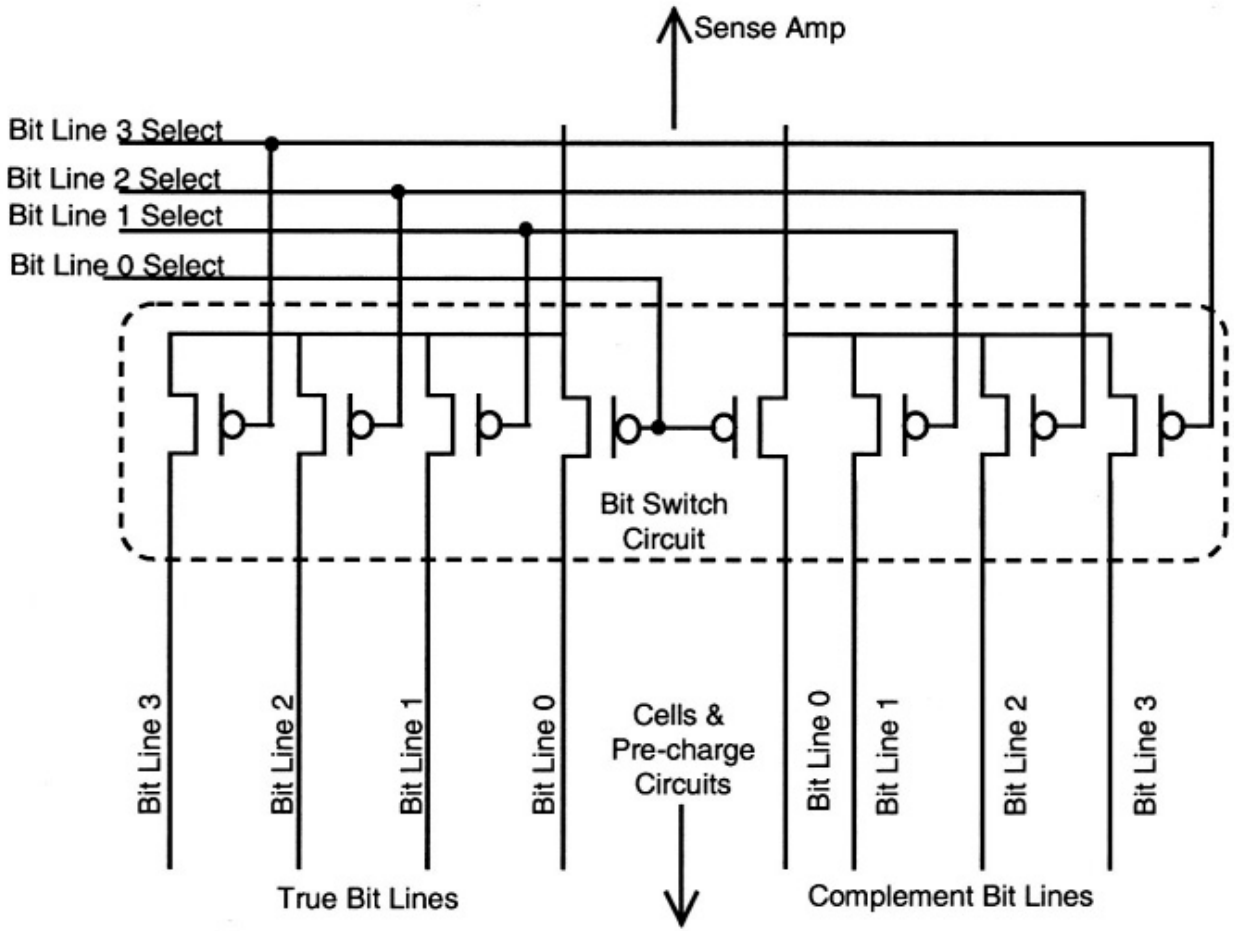
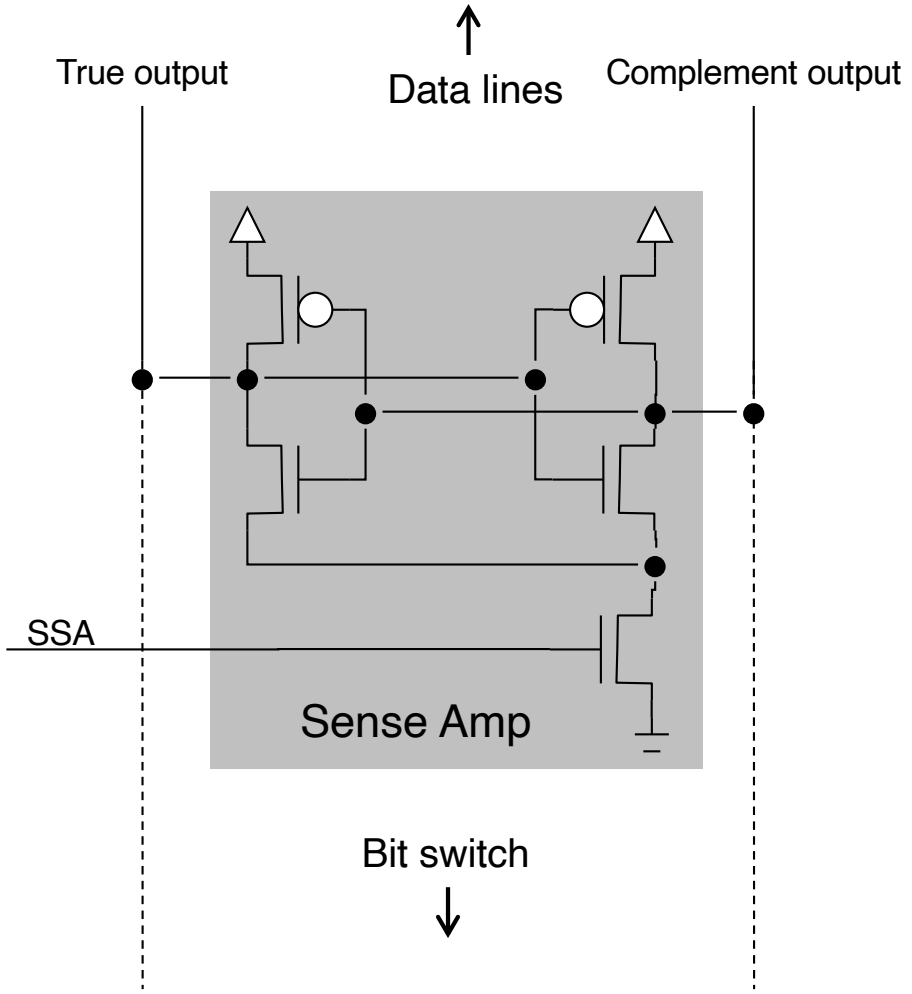


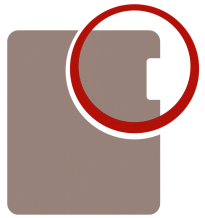
Figure 2-13. Bit switch isolation circuit.

Sense Amplifier

- Many types of sense circuitry. Shown here is a latch-type one.
- A second stage of sensing is often utilized to improve overall performance and latch the sensed result, regardless of the first stage's sense amp design style.



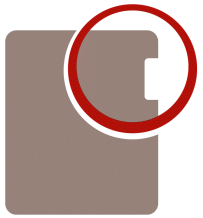
The ISO & SSA Control Signals



- ISO: bit line isolation
- SSA: set sense amp

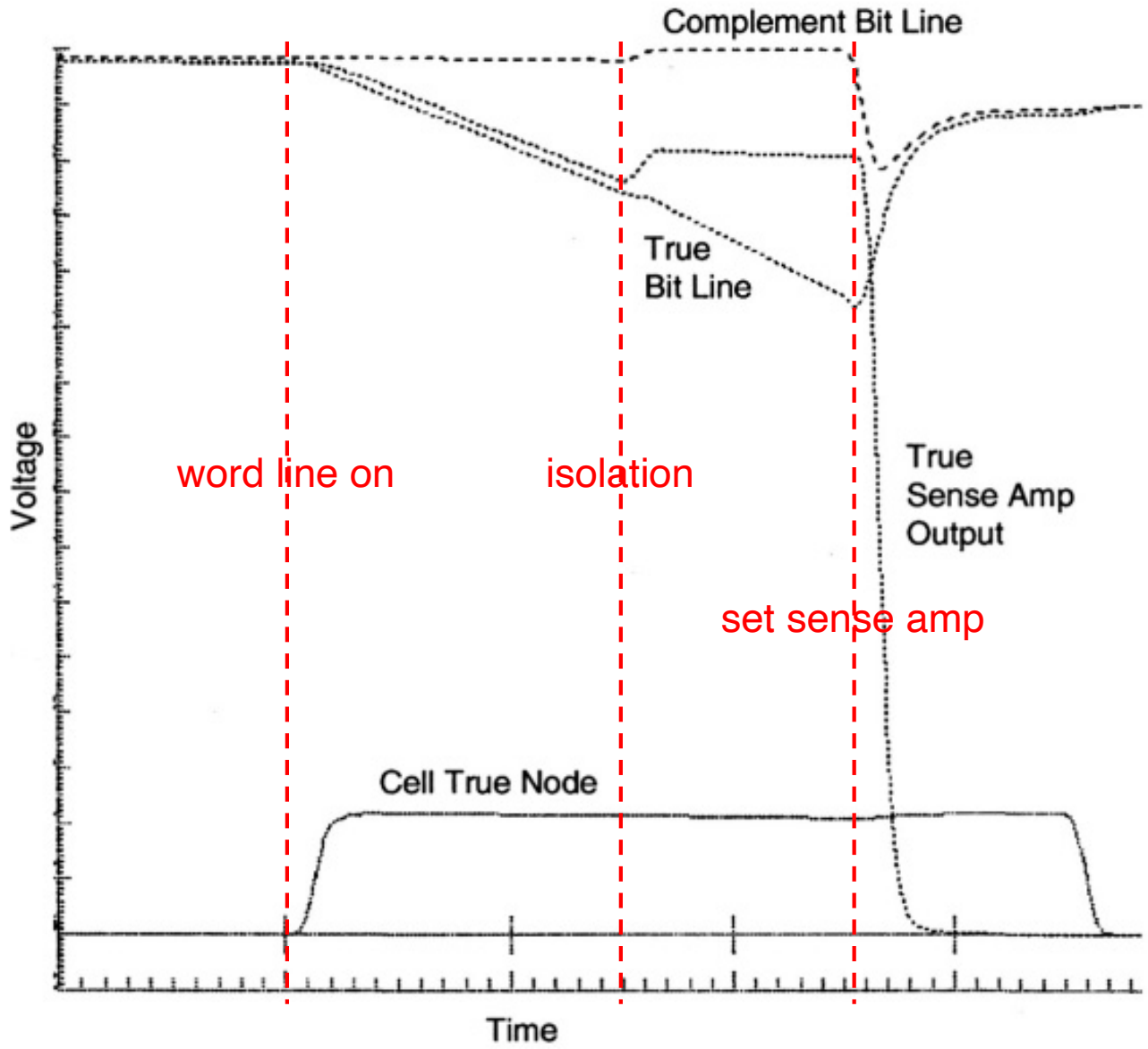
- The timing relationship between the ISO and SSA signals is configuration dependent.
  - One sense amp per bit line pair: can use SSA to drive ISO.
  - Shared sense amp through the bit switch: SSA slightly before ISO.

The SSA Signal



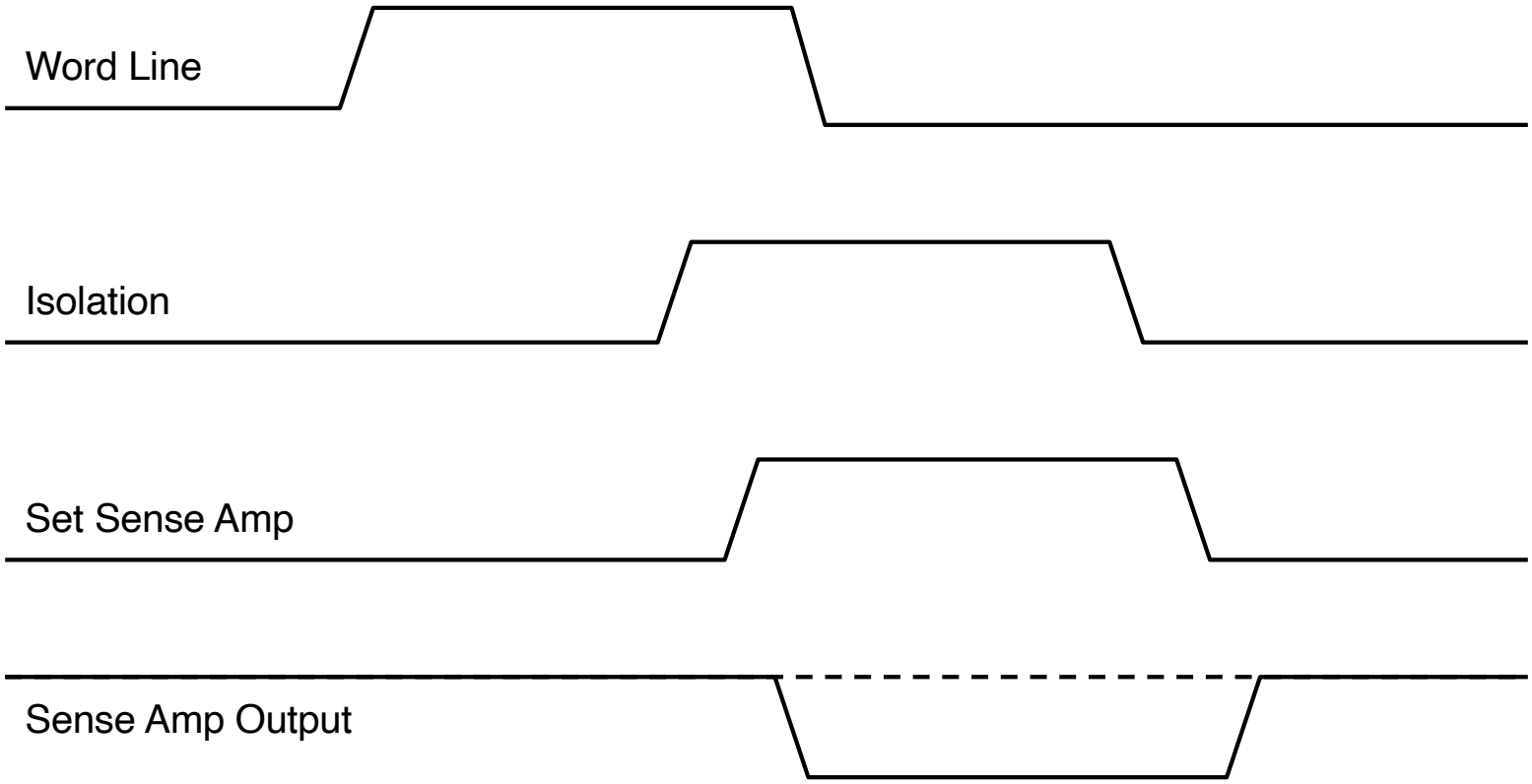
- The timing of SSA is critical.
  - Too late: waste of time.
  - Too early: cause insufficient signal to reach the sense amp.
- The SSA delay must be very accurately modeled and designed robustly to track the bit line signal development.
  - In the existence of process, voltage, temperature variations.
  - Some solutions: dummy word line, dummy bit line, …

An Example Read Timing (R0)

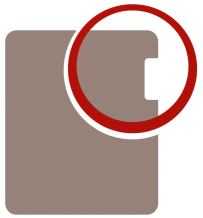


- In practice, ISO and SSA would be almost immediately adjacent in time.

Typical Read Timing

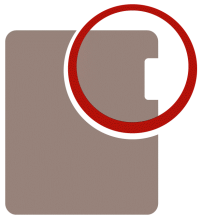


Remarks



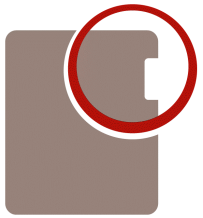
- There are numerous subtle design variations in the designs of memory cells, isolation devices, bit switches, and sense amps.
- Different designs may necessitate or eliminate the need of some fault models.

SRAM



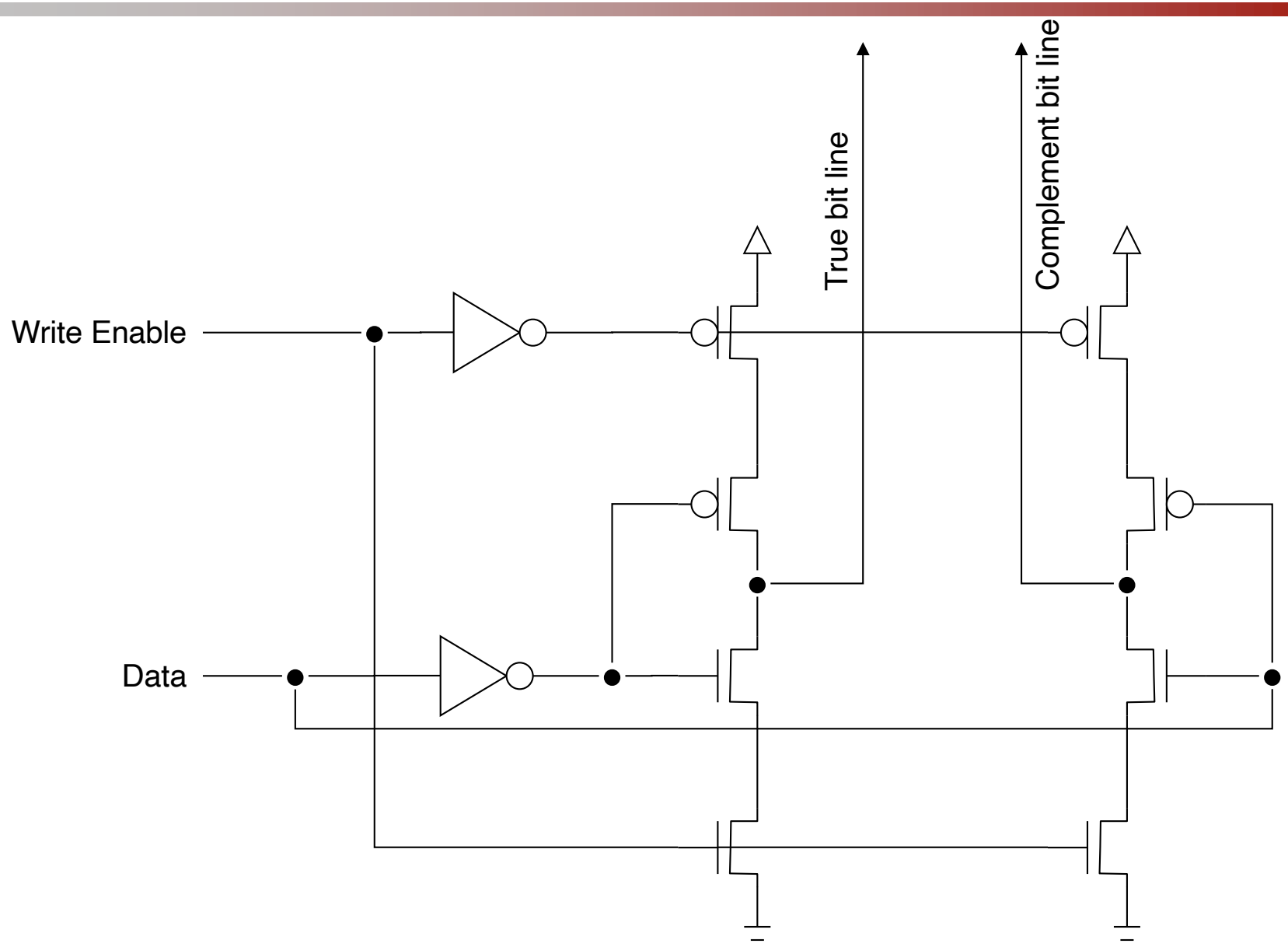
- SRAM cell design
- Read data path
- **Write driver circuit**
- Decoder circuitry
- Layout considerations
- Redundancy

The Write Operation

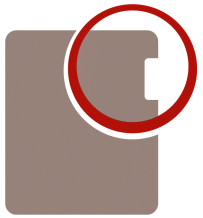


- Writing a cell is accomplished by writing a “0” into either the true or the complement side of the cell and the cell latch causes the opposite side to go to a “1” state.
  - The NFET transfer devices can drive a strong “0”but do not drive a “1” very effectively.

Write Driver Circuit

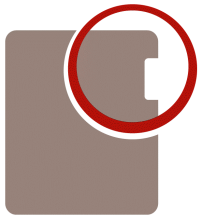


SRAM



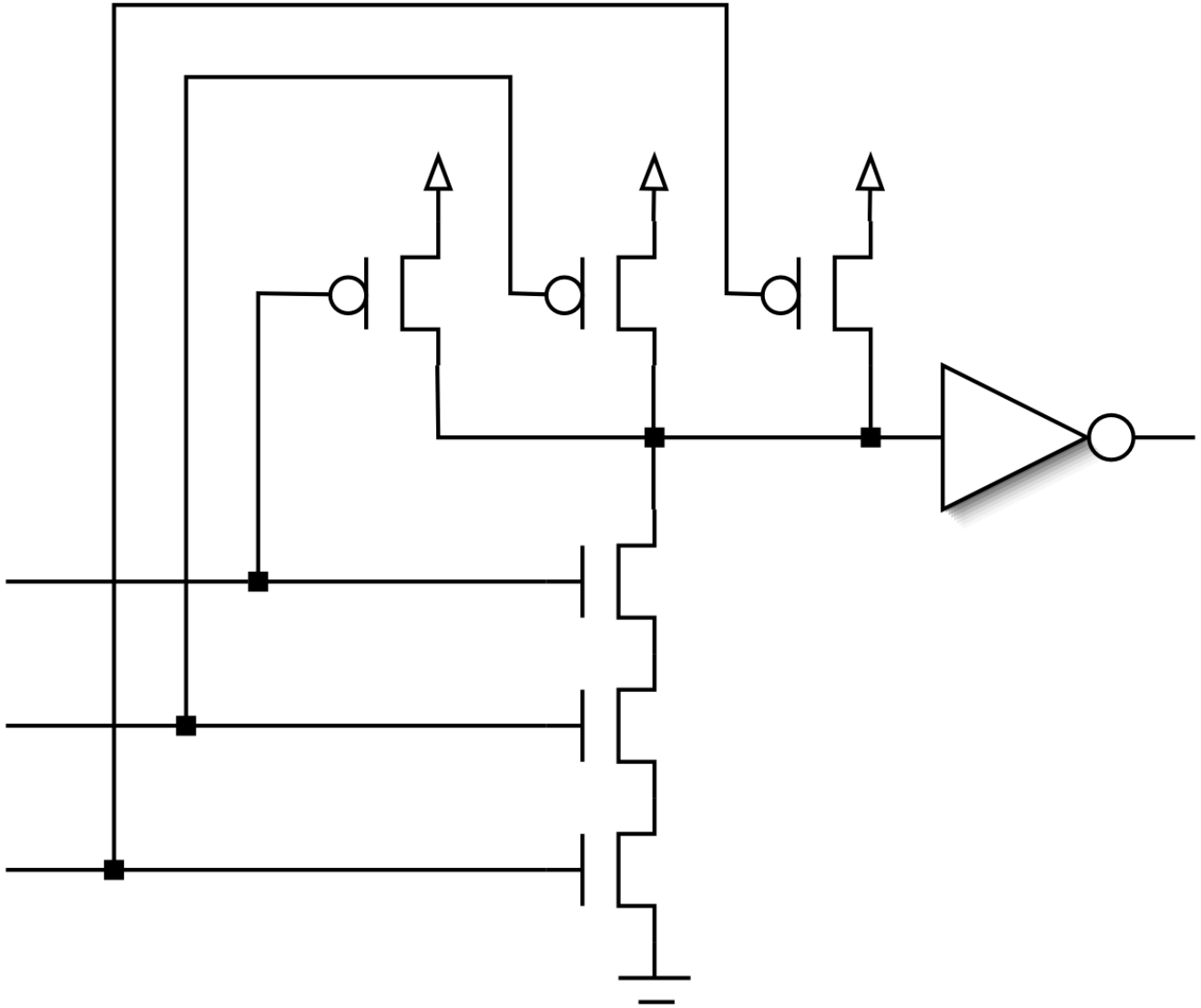
- SRAM cell design
- Read data path
- Write driver circuit
- **Decoder circuitry**
- Layout considerations
- Redundancy

Address

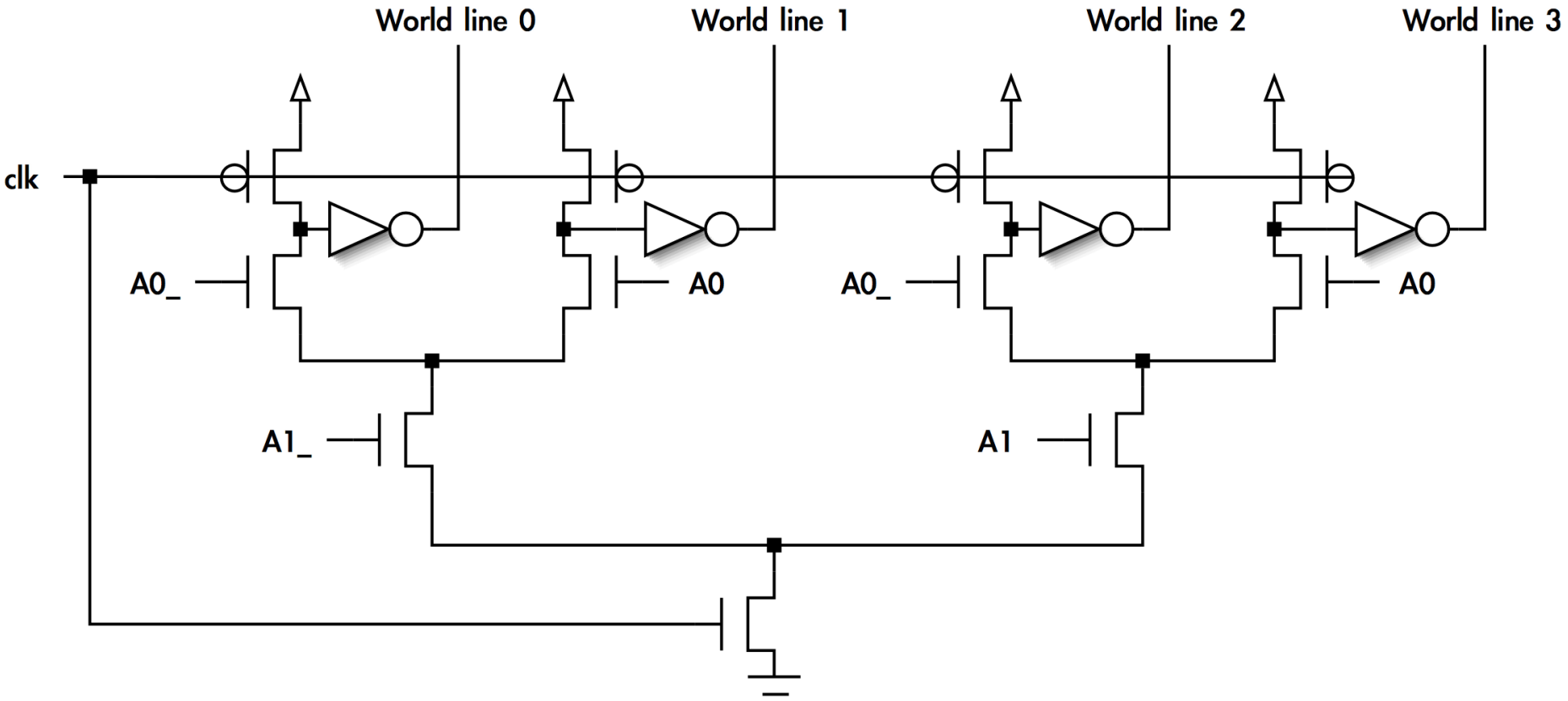


- For larger memories, the address is divided into row, column, and even bank addresses.
  - Row: the y coordinate.
  - Column: the x coordinate.
  - Bank: sub-array, quadrant, or other terms.
- Different types of decoding circuitry may be employed for different addressing.

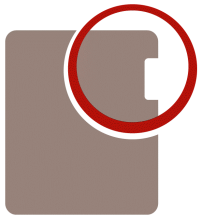
Example Static Decoder Circuitry



Example Dynamic Decoder Circuitry

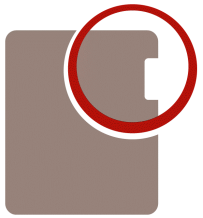


SRAM



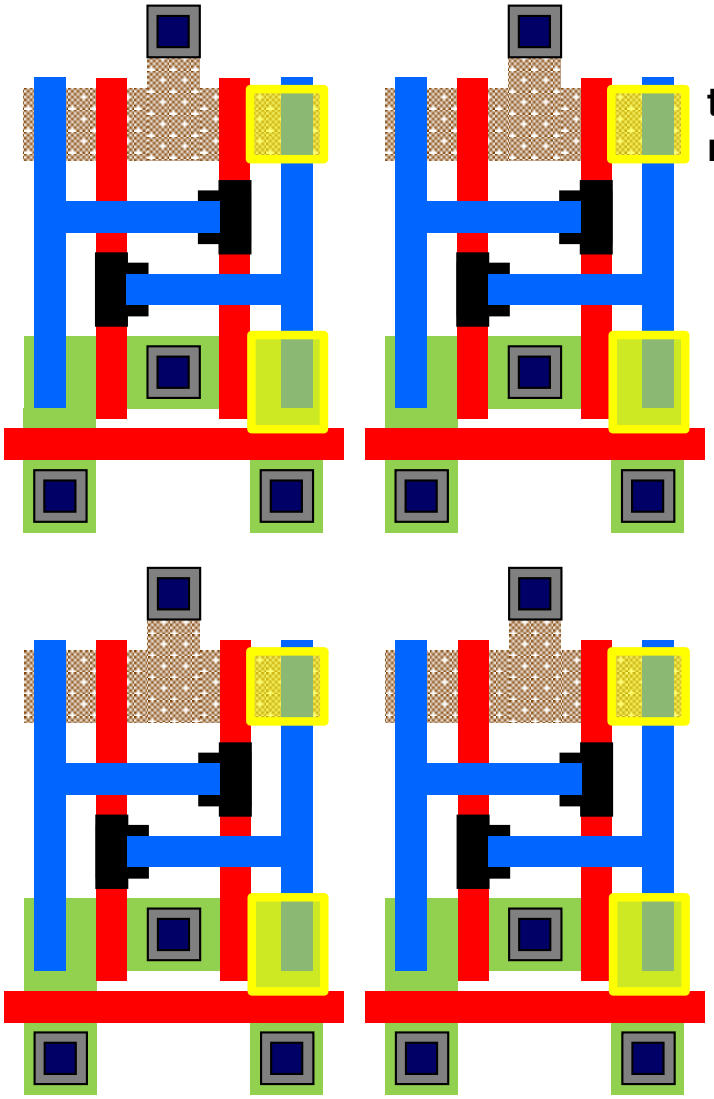
- SRAM cell design
- Read data path
- Write driver circuit
- Decoder circuitry
- **Layout considerations**
- Redundancy

Layout Considerations

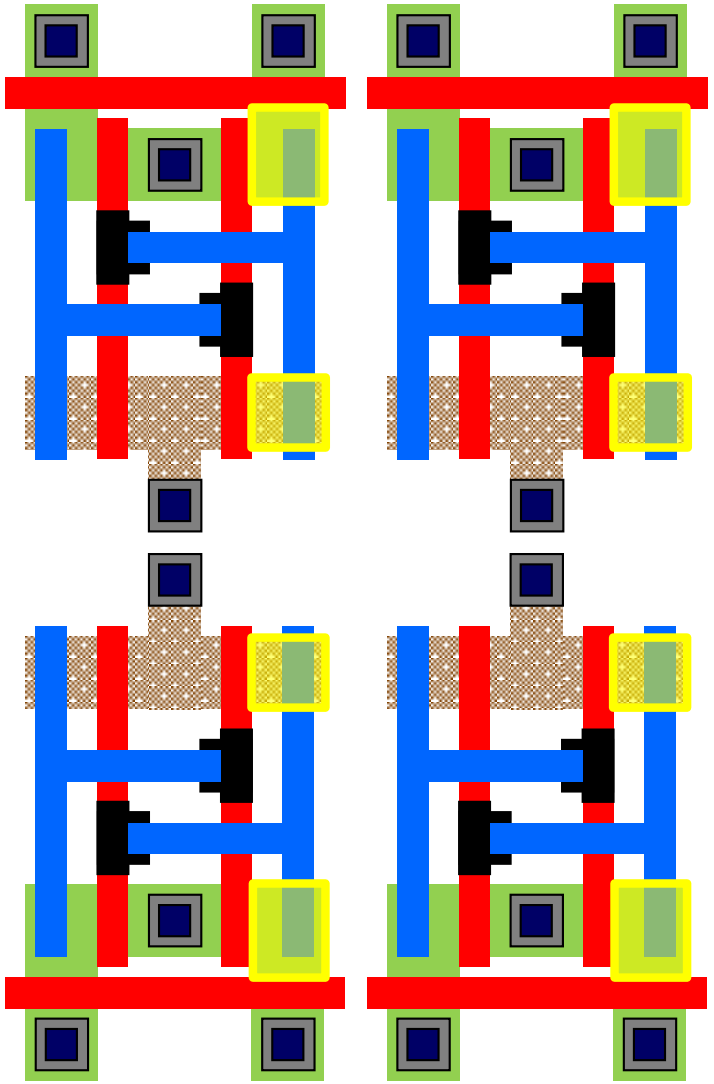


- Special attentions are paid to memory layout for various reasons:
  - Performing cell stepping, mirroring, and rotating to optimize overall memory size. (This can also be done in a sub-array basis.)
  - Spread bits of a single word in multiple sub-arrays for protection against soft errors.
  - Twisted bit lines to reduce capacitance coupling between adjacent bit lines.

Cell Stepping, Mirroring, and Rotating

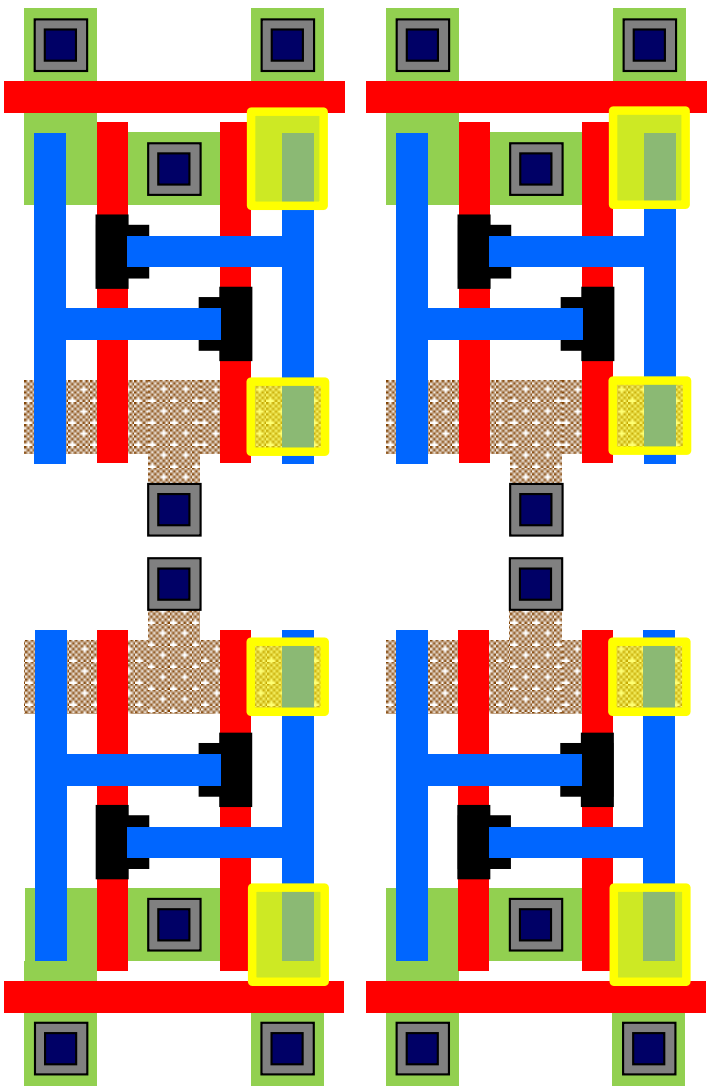


Four stepped cells

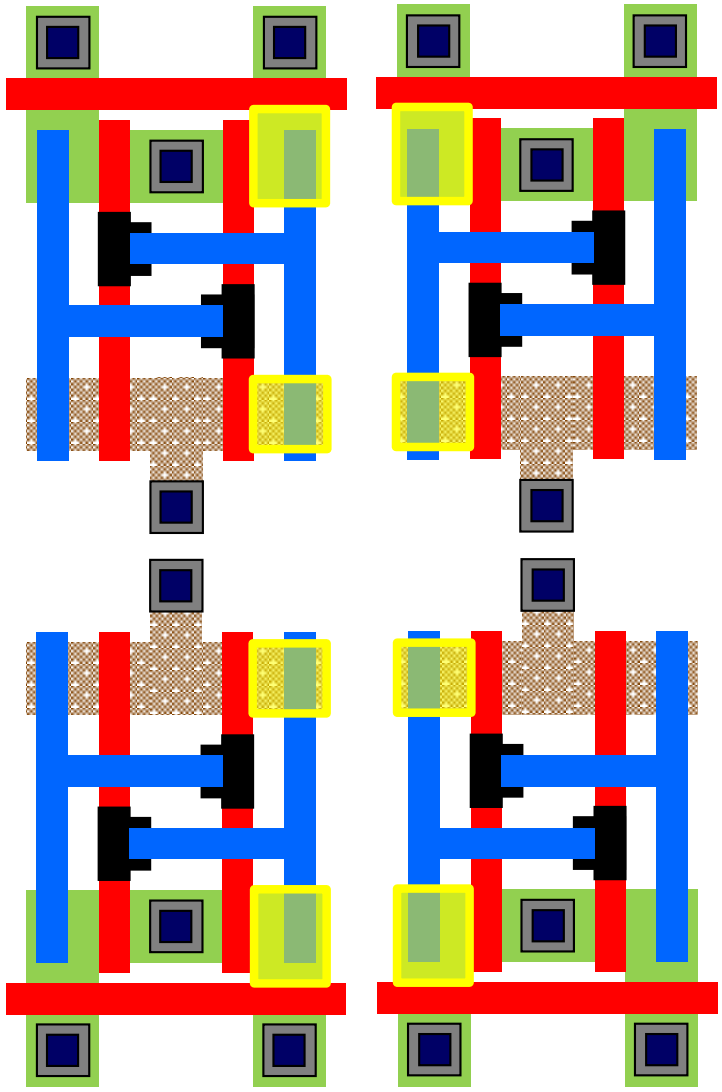


Mirrored about the x-axis

More ...

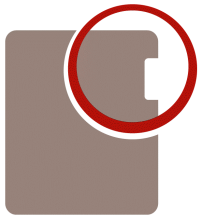


Mirrored about the x-axis



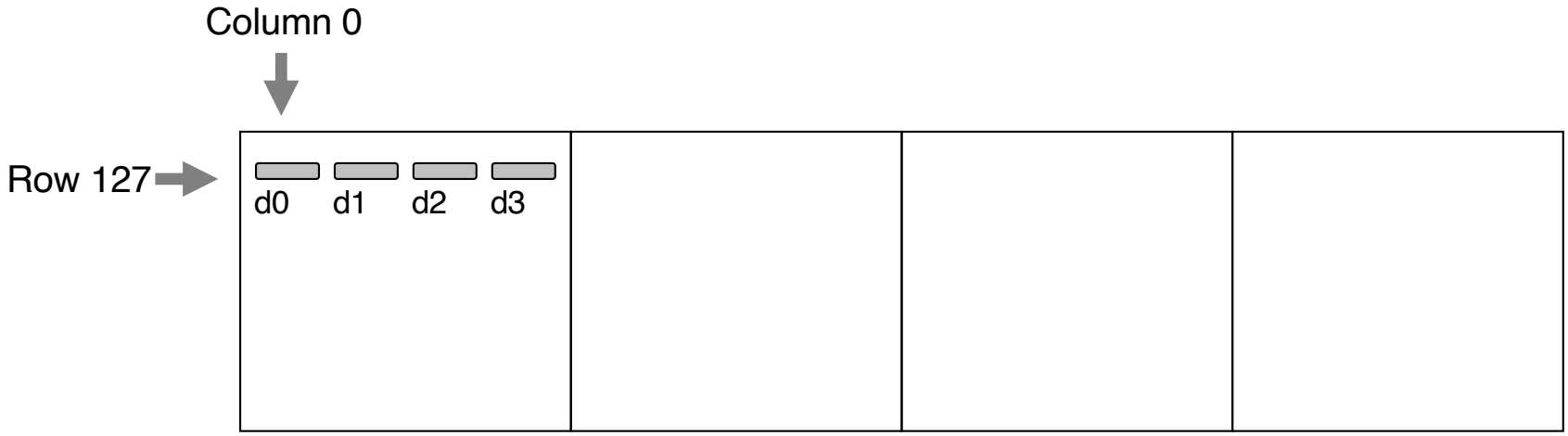
Mirrored about the x and y-axis

Word Storage

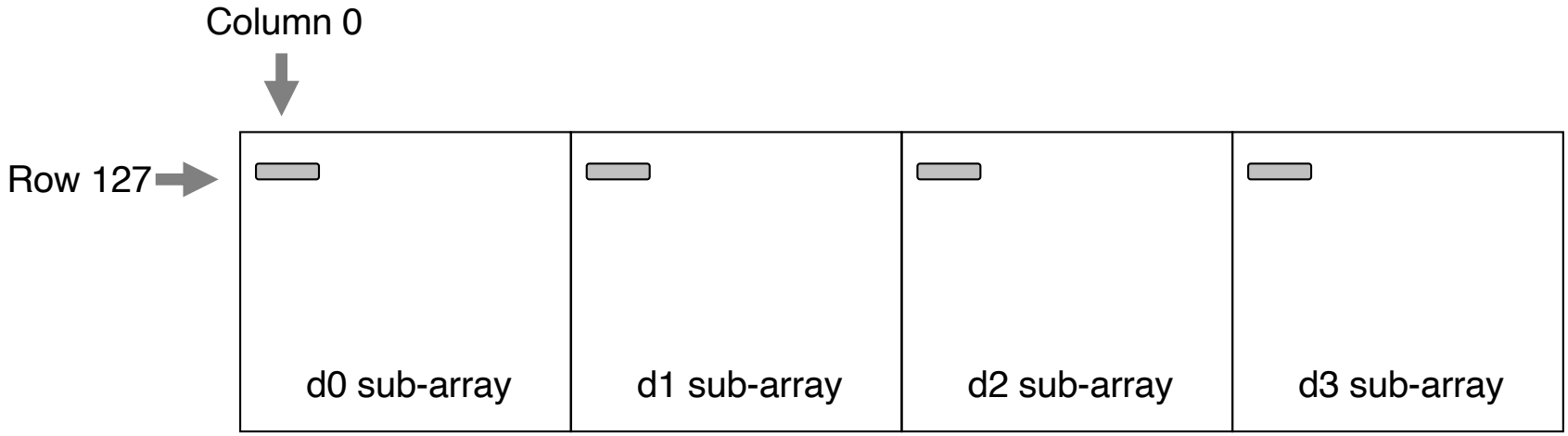


- All of today's memories have more than a single I/O.
  - An address points to a "word," not a single bit.
- Usually, the cells that store the bits from a single word are spread across multiple sub-arrays.
  - One of the reasons is protection against soft errors by reducing the possibility of two bits flip which cannot be detected by parity check.

Word Storage - cont'd

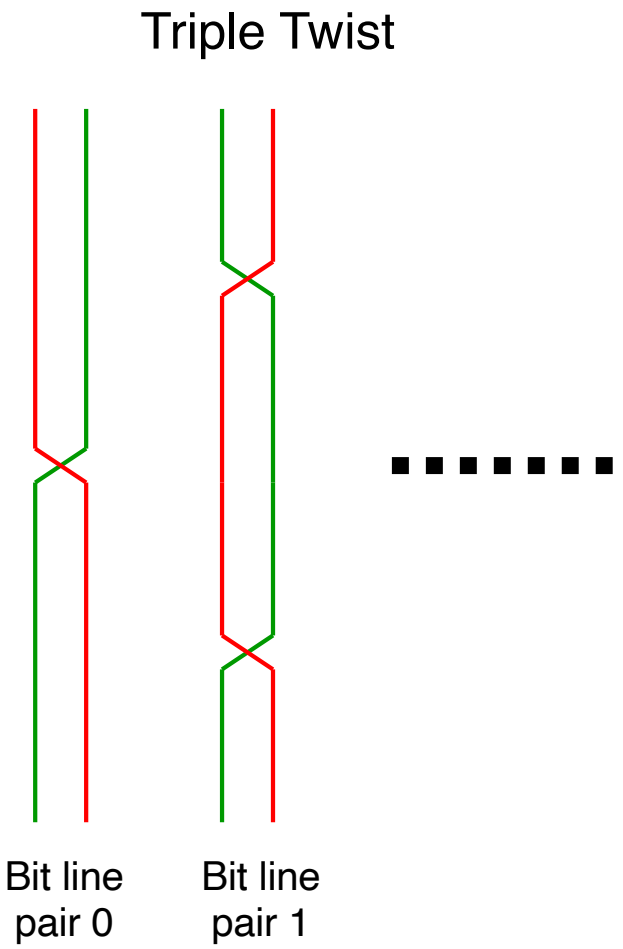


Entire word in a single sub-array



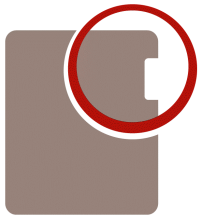
Word spreads across multiple sub-arrays

Bit Line Twisting



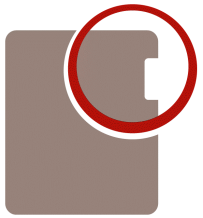
- To reduce the capacitance coupling between long adjacent bit lines.
- Triple twist:
  - 1st pair twisted at half point.
  - 2nd pair twisted at 1 and 3 quarter points.
  - …
- The triple twist configuration will force any coupling to be common mode which makes the sense amp less prone to erroneous evaluations.

SRAM



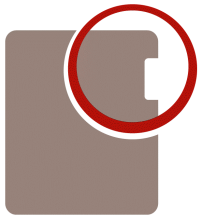
- SRAM cell design
- Read data path
- Write driver circuit
- Decoder circuitry
- Layout considerations
- **Redundancy**

Redundancy



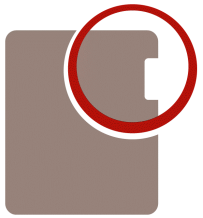
- Memories require redundancy to ensure that sufficient chip yield is obtained.
- A redundant element is a piece of memory that can replace a defective piece of memory.
  - In the form of spare rows, I/O, columns, blocks, or a combination of the above.
- The need for redundancy decreases on a per bit basis, but the total amount of redundancy needed is growing on a per chip basis.
- We will discuss this matter in later units.

Outline



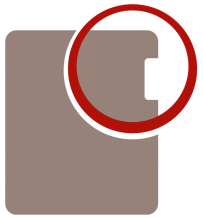
- Overview
- Design and test considerations
  - SRAM
  - **Multi-port memories**
  - DRAM
- Memory testing
- Memory self test

Multi-Port Memories



- Allows more than one operation per cycle.
  - Useful when a designer would like to read data in each cycle and would also like to write data in each cycle.
- The number of multi-port memory applications will continue to grow as system engineers strive to increase performance and recognize the availability and utility of multi-port memories.

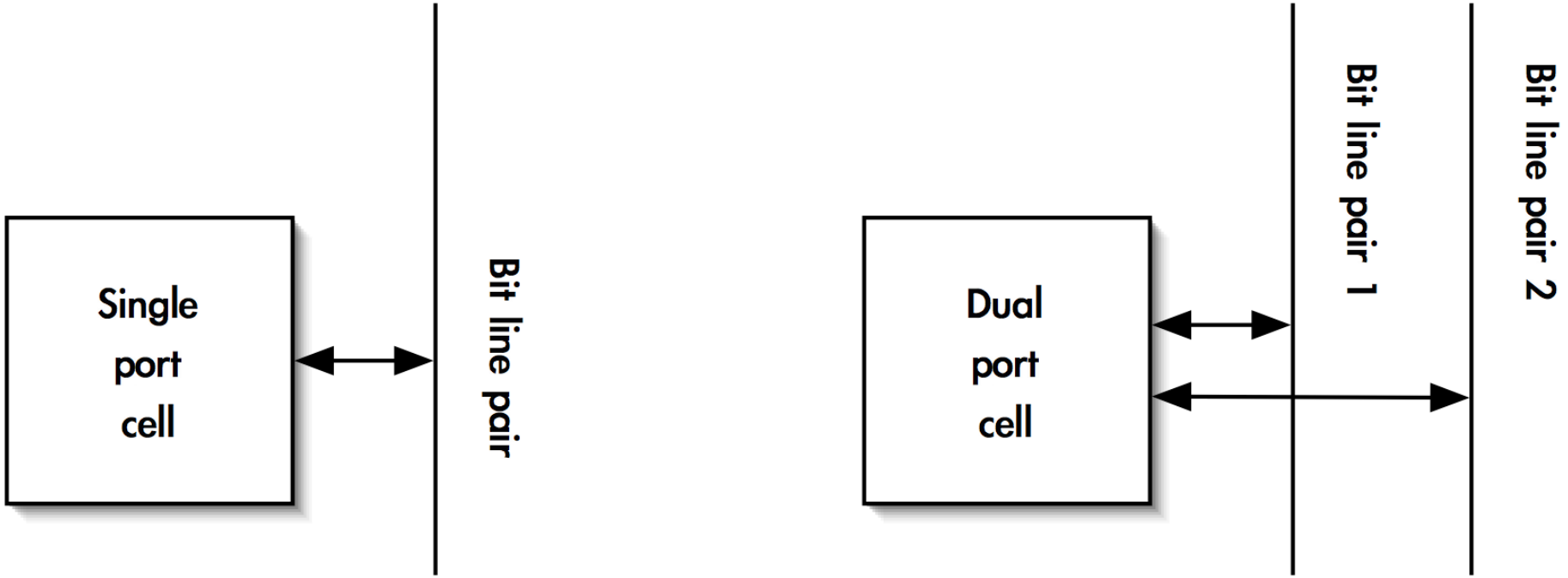
Multi-Port Memories



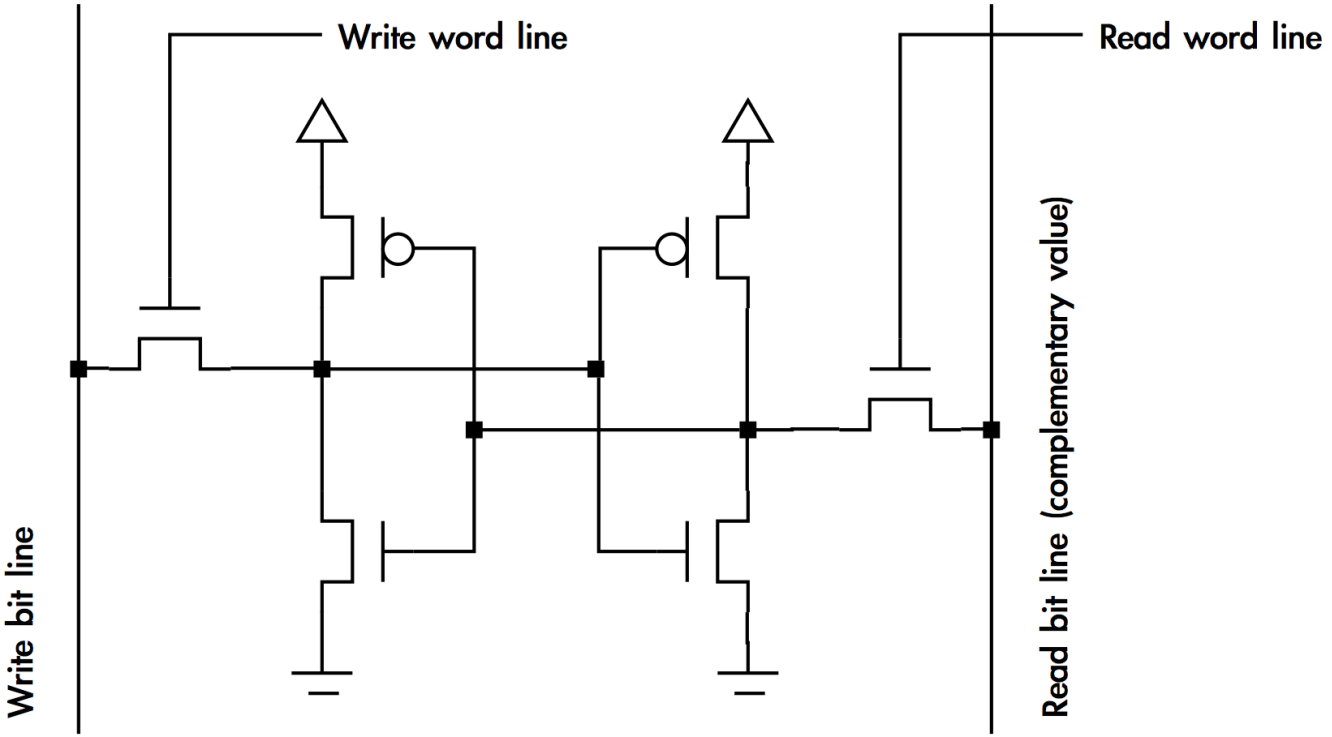
- **Cell basics**
- Timing issues
- Layout considerations

Block Diagram

- A multi-port memory cell has to have more than one access port.
- We assume multi-port SRAM here, although it is possible to have other types of multi-port memories.

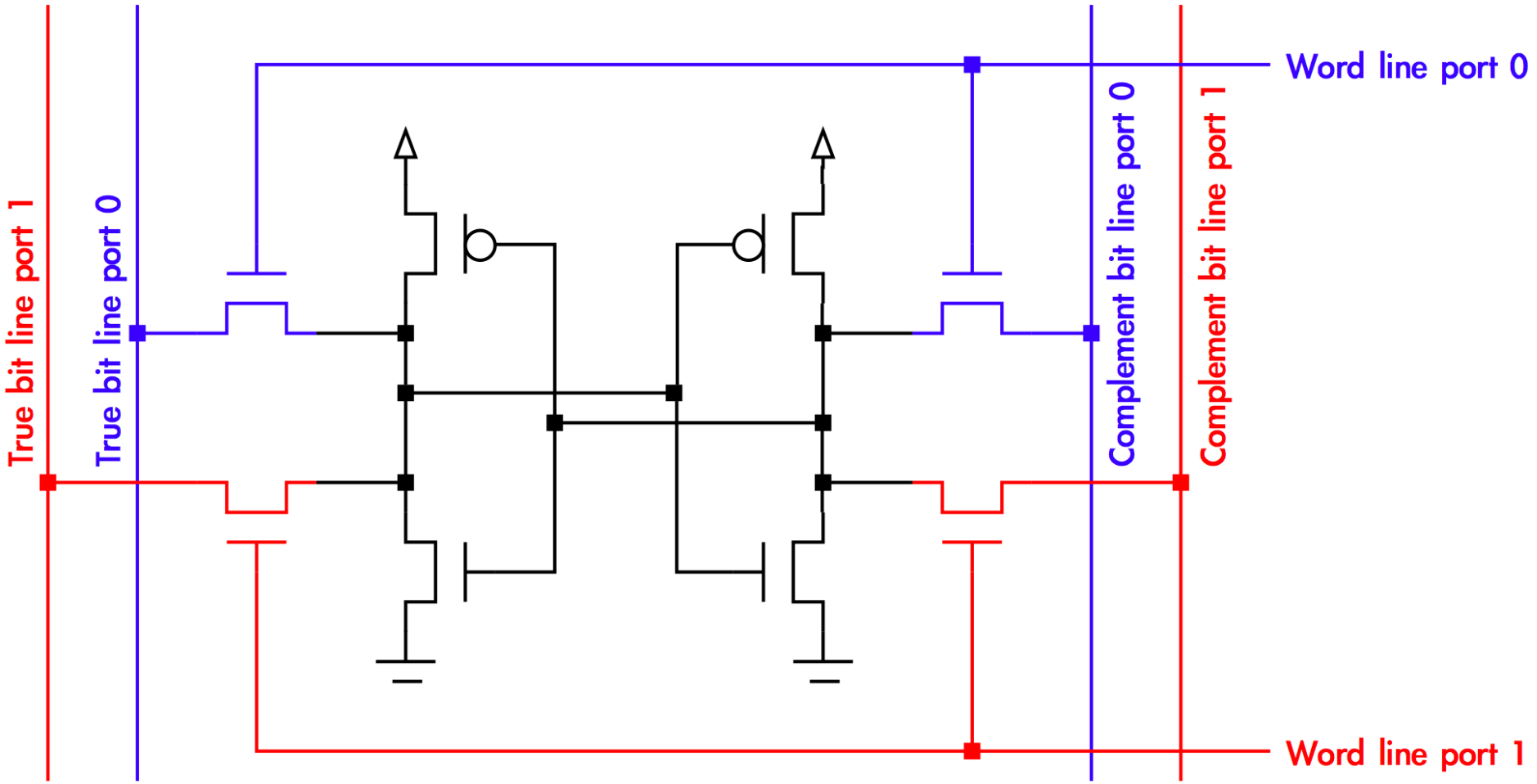


A Simple Two-Port Cell

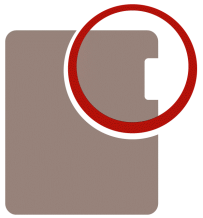


- Single-ended read operation is slow.
  - Need larger voltage swing, but simple sensing scheme, e.g., an inverter.
- NFET is not very effective at writing 1.
- This topology should be avoided.

The Eight-Device Two-Port Read/Write Cell



Stability of the Multi-Port Cell

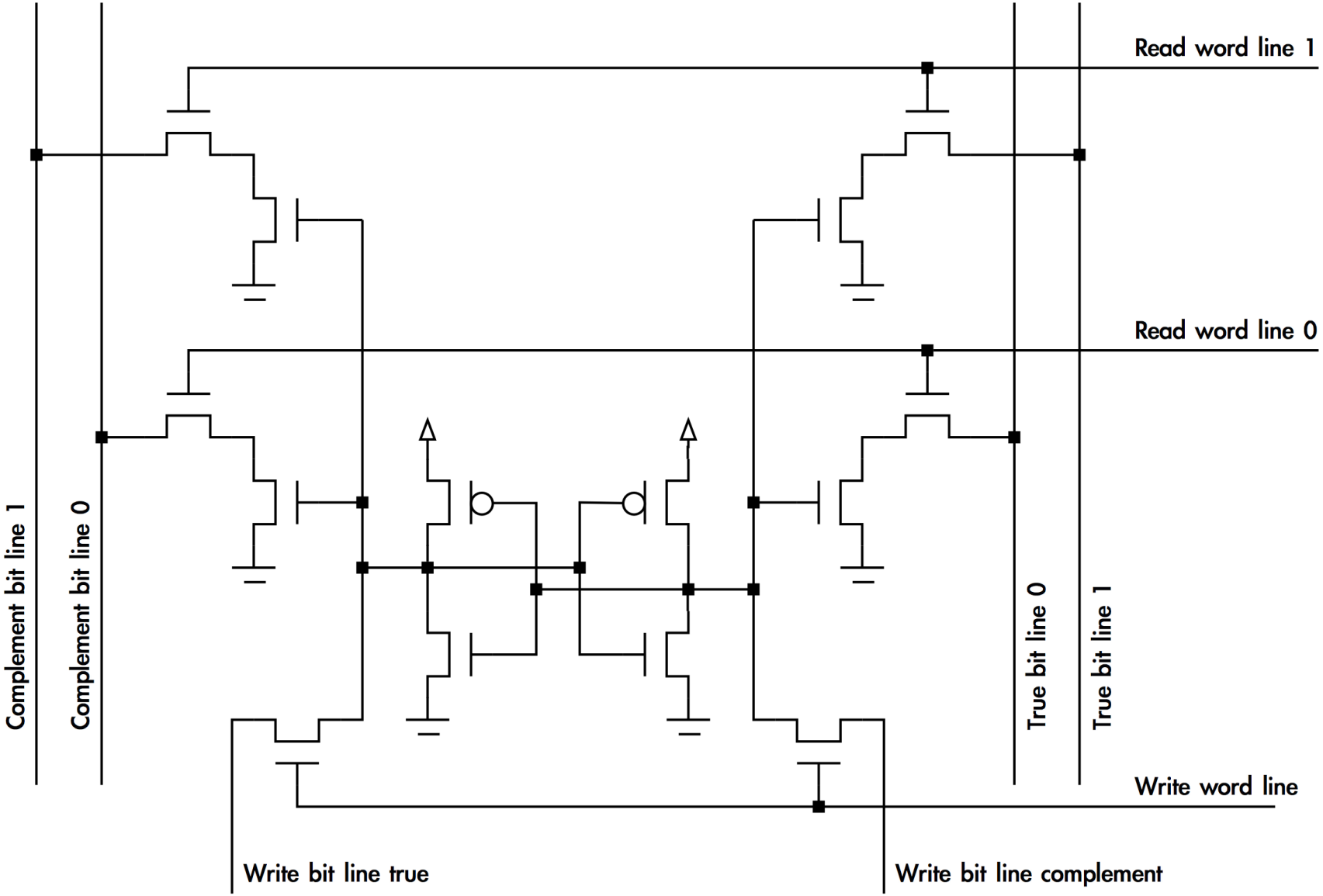


- Much trickier than a one-port cell.
  - Simultaneous operations are performed regularly.
- The beta ratio for an *n* read port memory.

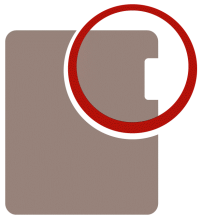
$$\beta = \frac{W_{PullDownFET} / L_{PullDownFET}}{\sum_{n=1}^{\#ReadPorts} \left(W_{TsfrFET_n} / L_{TsfrFET_n} \right)}$$

  - Typically, 2.0 to maintain cell stability.
  - Cannot be too stable to write.

Active Pull Down

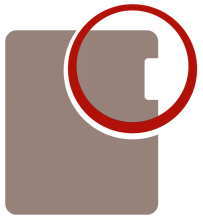


Multi-Port Memories

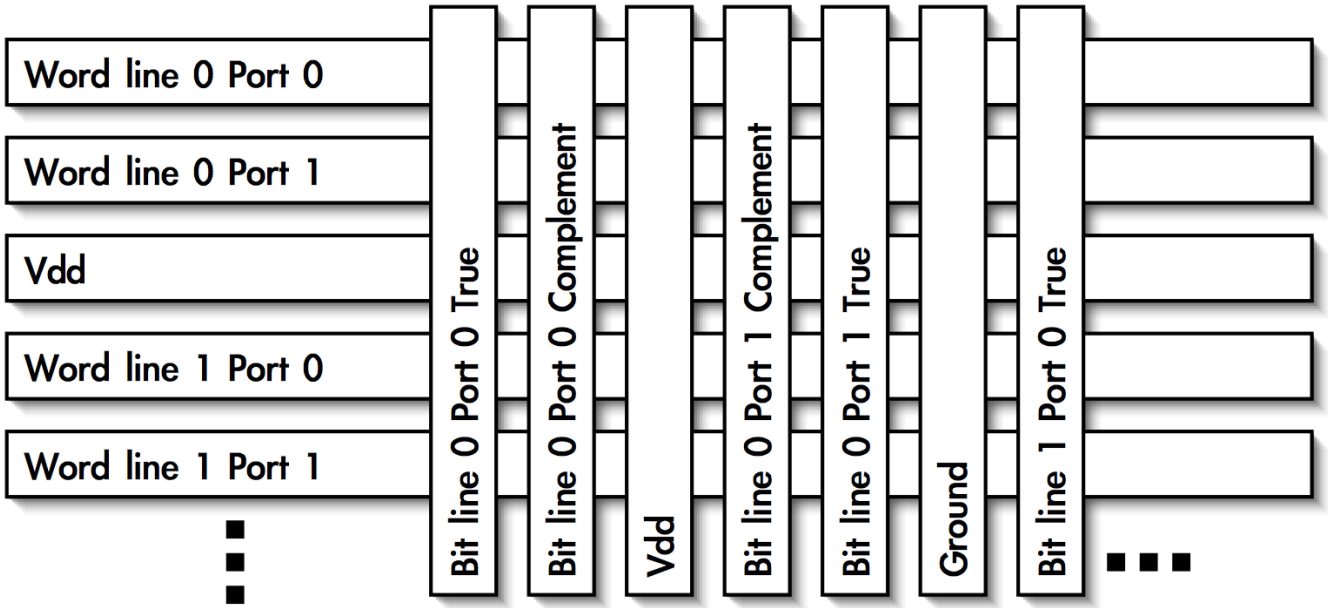


- Cell basics
- Timing issues
- **Layout considerations**

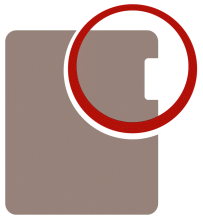
Capacitive Coupling



- Capacitive coupling between adjacent word or bit lines is a severe problem.
- Shielding needed to prevent excessive coupling into the read bit lines.
  - The read bit line is only held high capacitively.
  - Usually, alternating ground and Vdd lines.

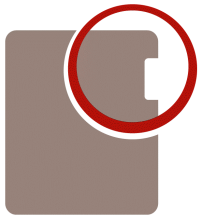


Other Considerations



- A multi-port memory cell is asymmetric.
  - Usually, only stepping is possible.
- Bit line twisting can become very difficult.
  - The coupling concerns must be examined more carefully.
- Redundancy may become a necessity as the multi-port memory size grows.
  - A failure, on any port, must cause complete replacement of the cell with a redundant one.

Summary



- Multi-port memories have many similarities to standard one-port SRAMs.
- The possible interactions between the ports must be carefully analyzed.