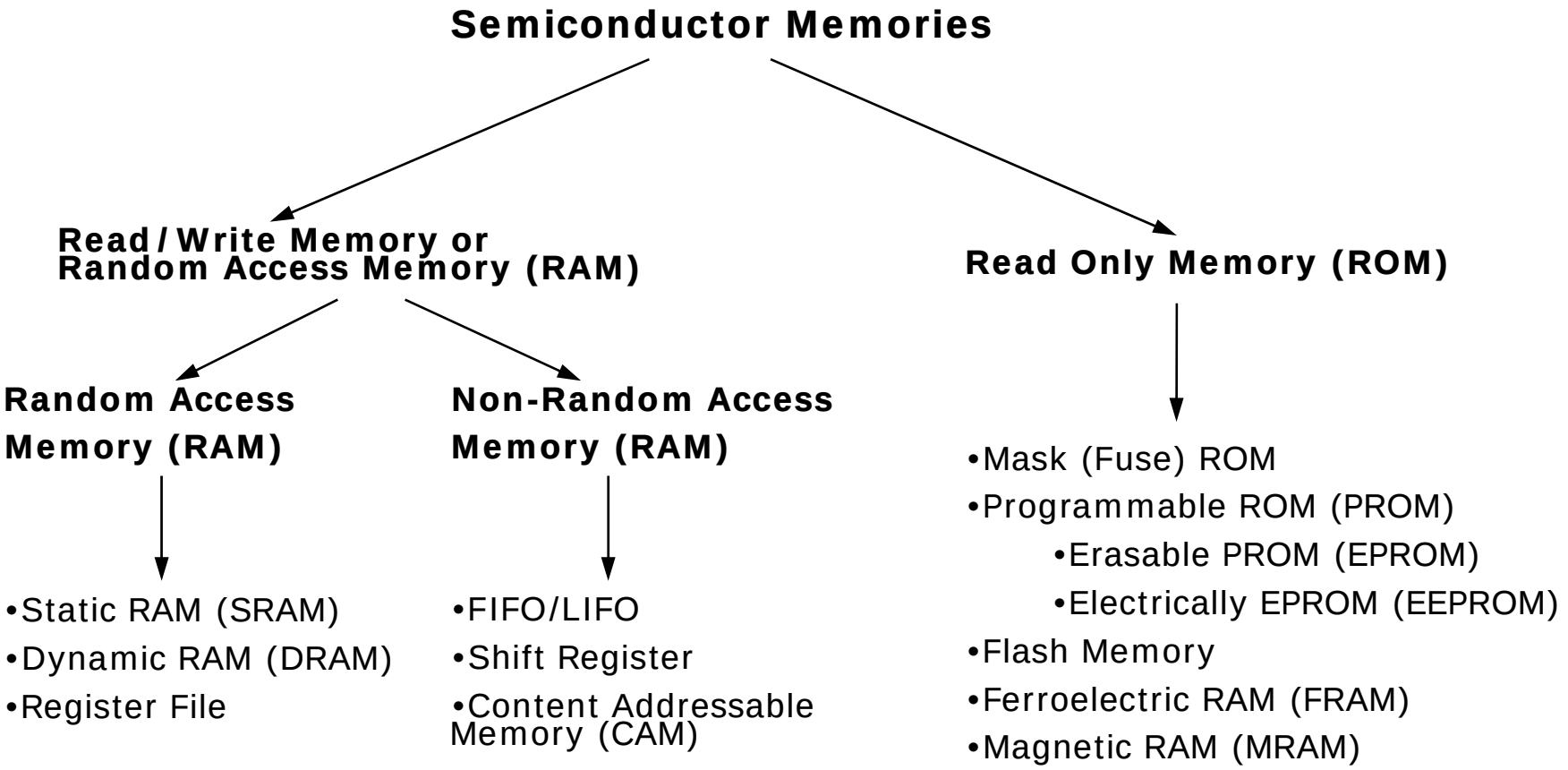
Chapter 3
Semiconductor Memories

Jin-Fu Li
Department of Electrical Engineering
National Central University
Jhongli, Taiwan

Outline

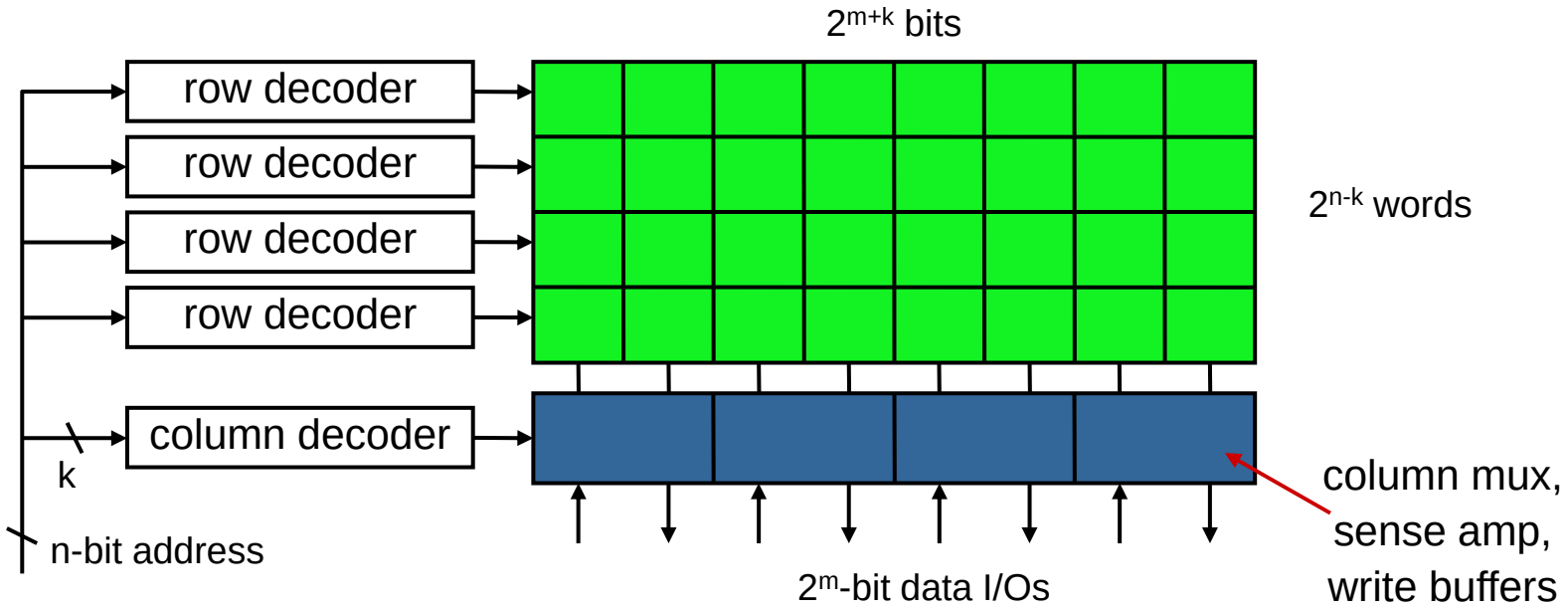
- Introduction
- Random Access Memories
- Content Addressable Memories
- Read Only Memories
- Flash Memories

Overview of Memory Types

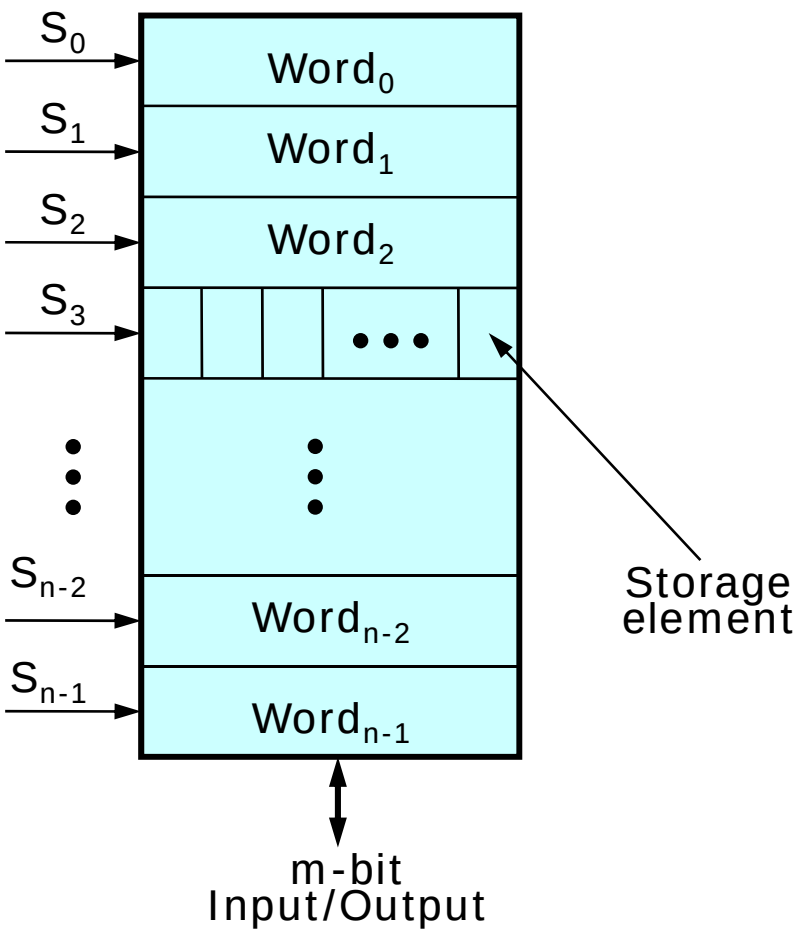


Memory Elements – *Memory Architecture*

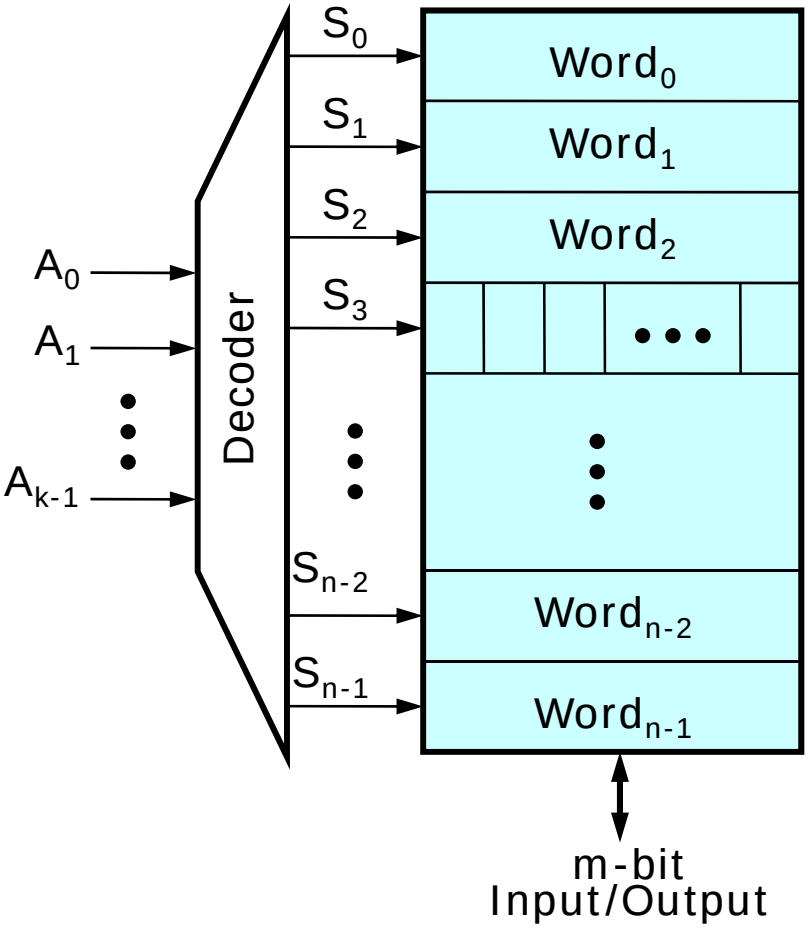
- Memory elements may be divided into the following categories
  - Random access memory
  - Serial access memory
  - Content addressable memory
- Memory architecture



1-D Memory Architecture

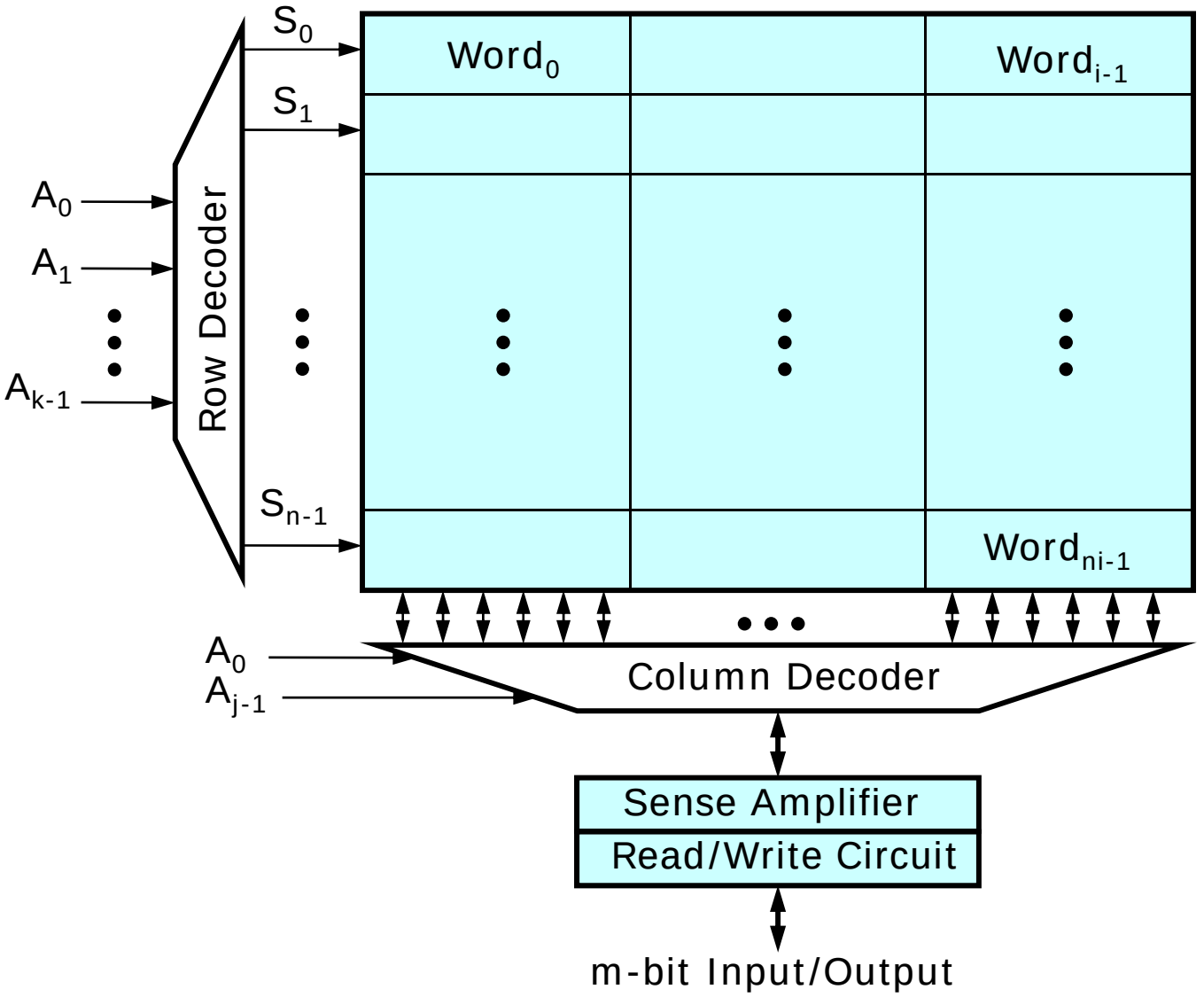


n select signals: S_0-S_{n-1}

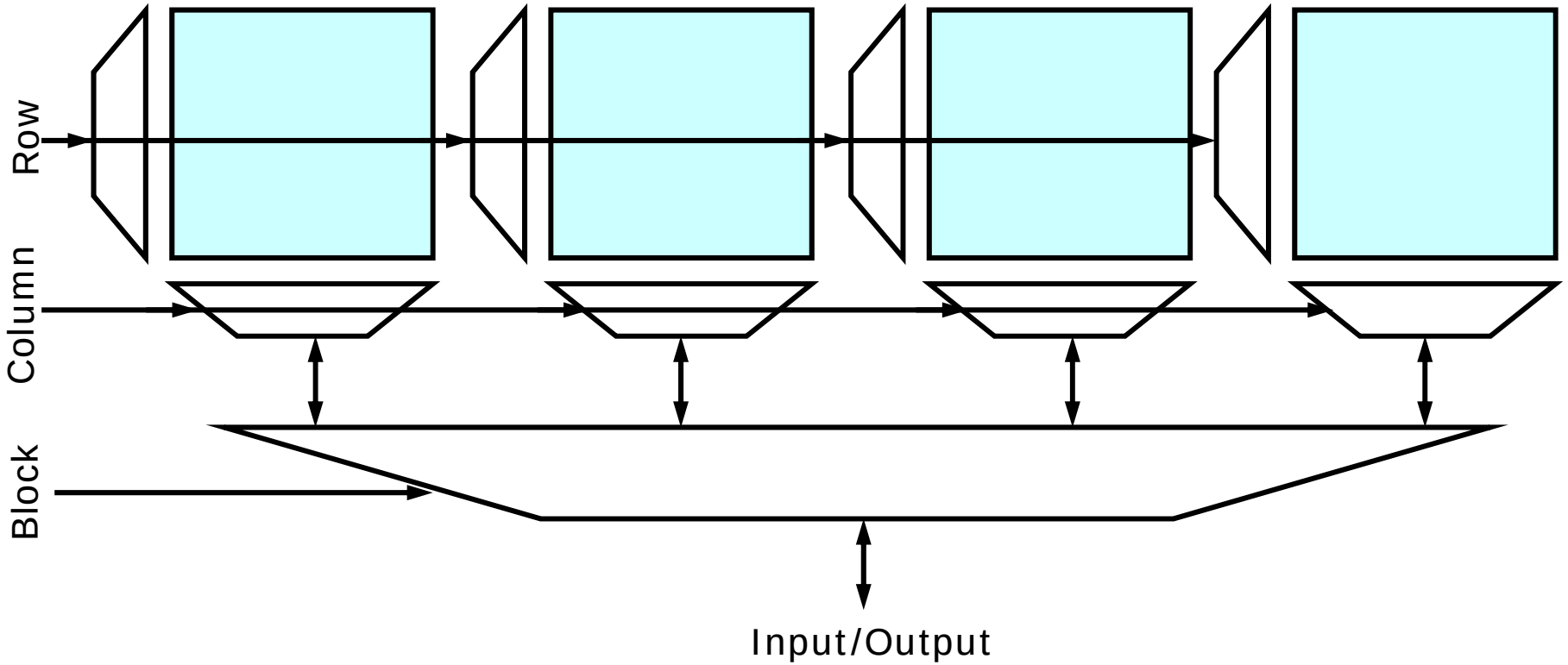


n select signals are reduced to k address signals: A_0-A_{k-1}

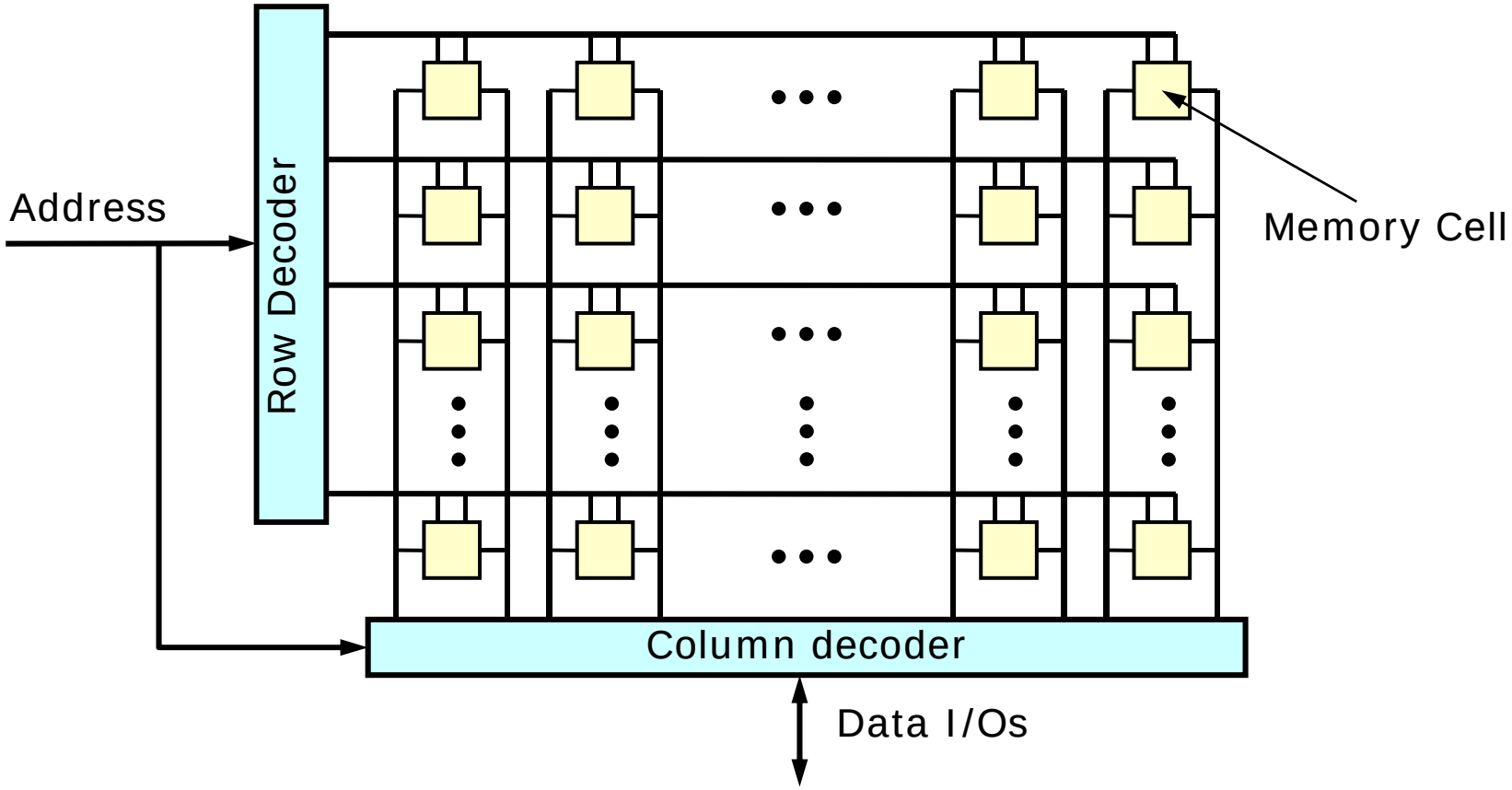
2-D Memory Architecture



3-D Memory Architecture

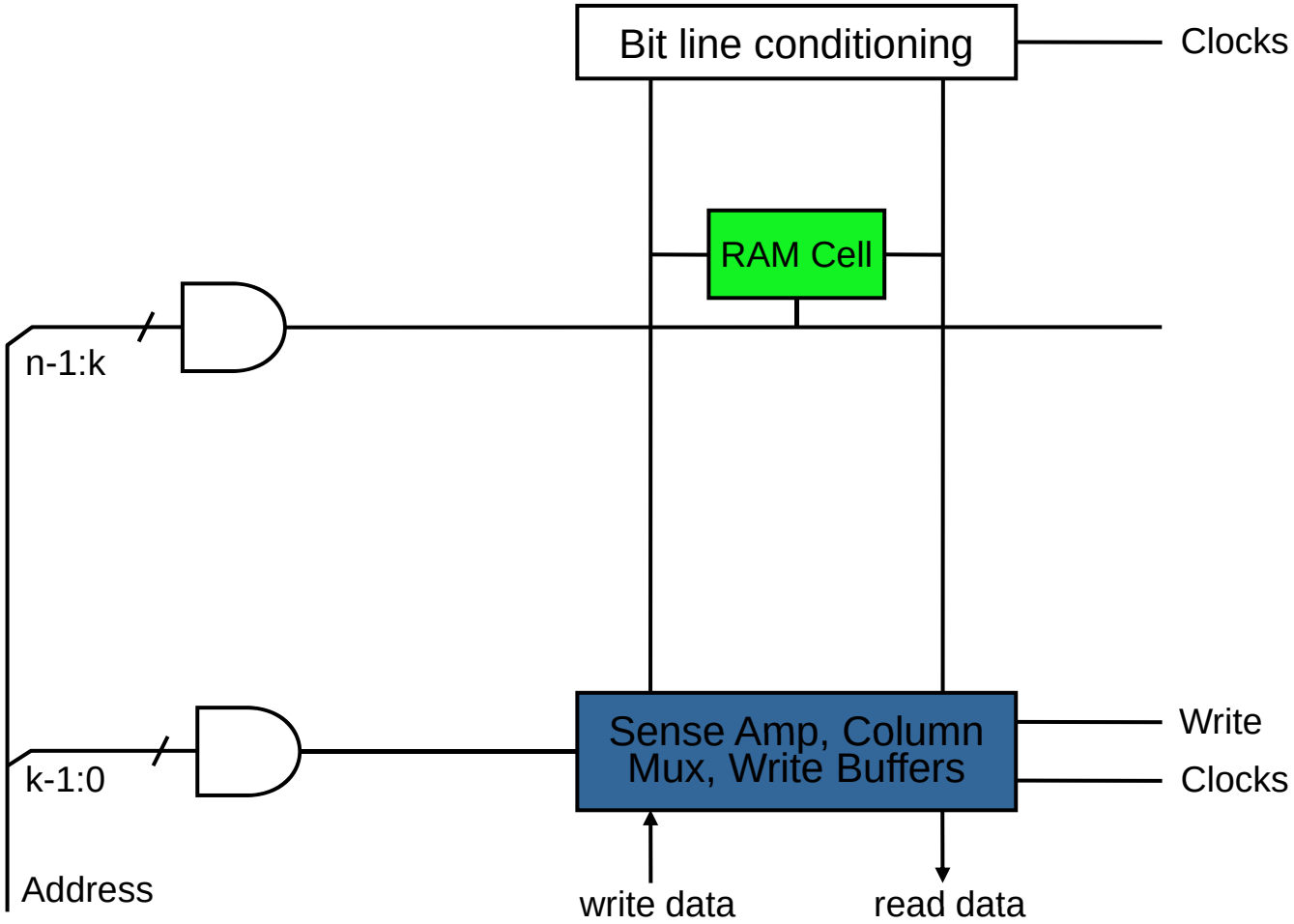


Conceptual 2-D Memory Organization



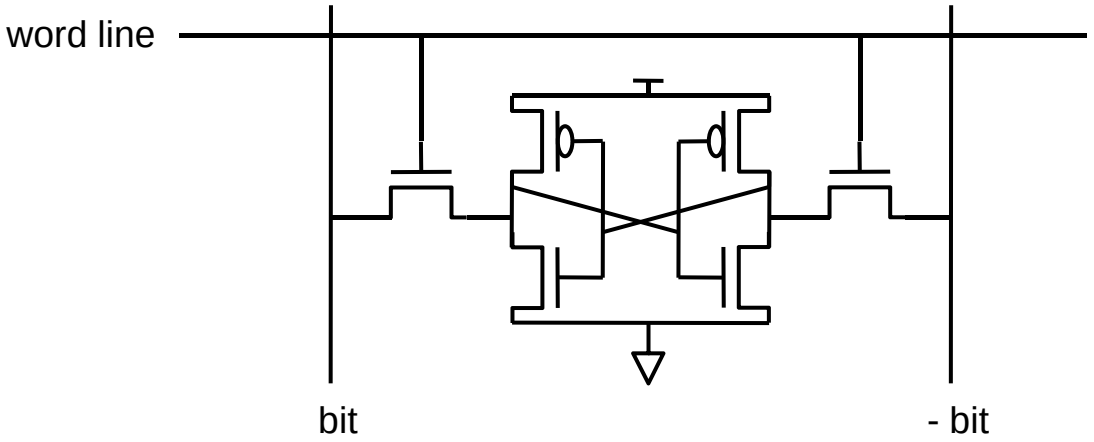
Memory Elements – *RAM*

Generic RAM circuit

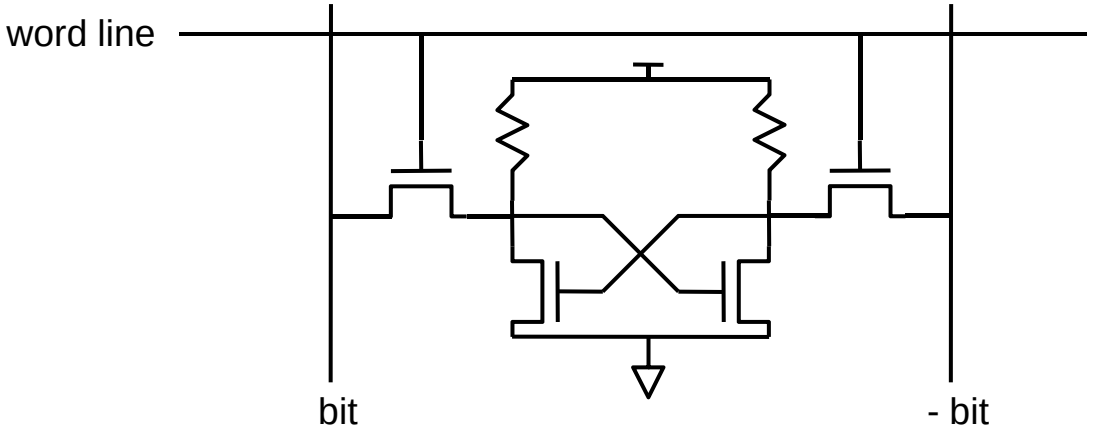


RAM – *SRAM Cells*

6-T SRAM cell

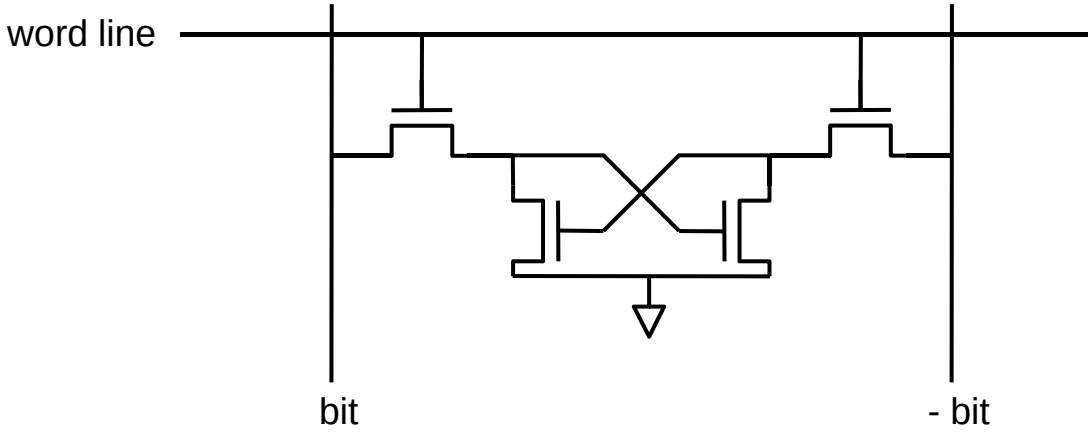


4-T SRAM cell

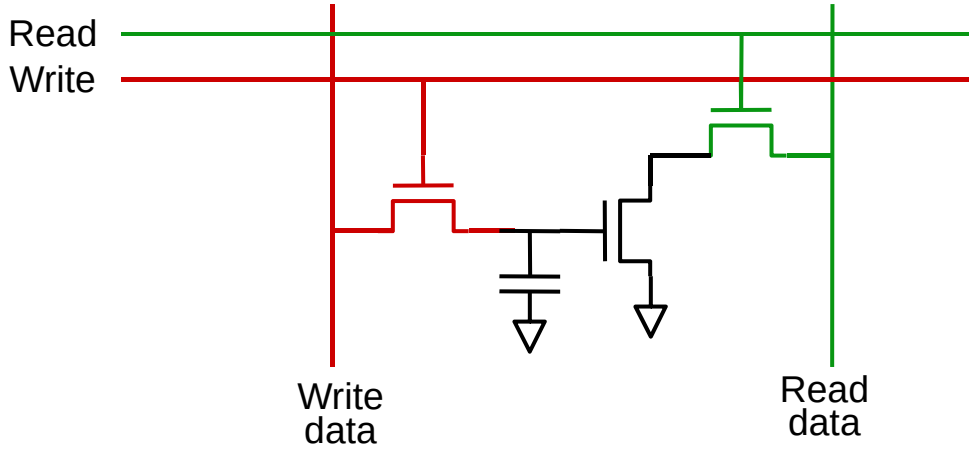


RAM – *DRAM Cells*

4-T dynamic RAM (DRAM) cell

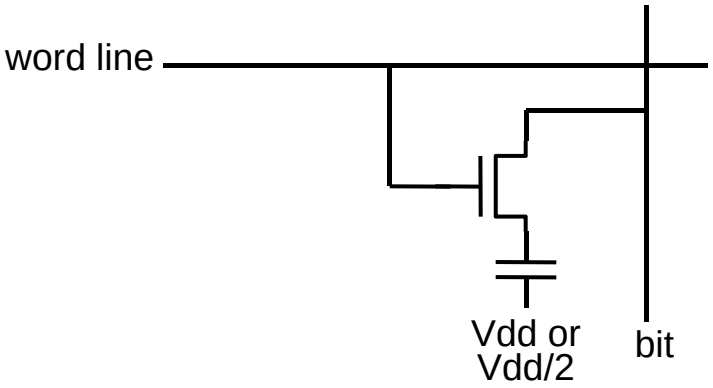


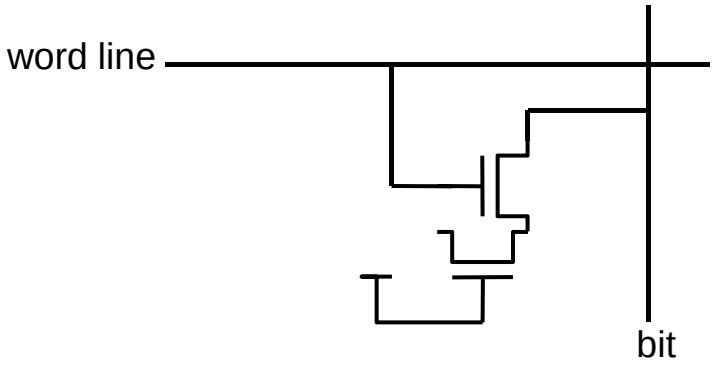
3-T DRAM cell



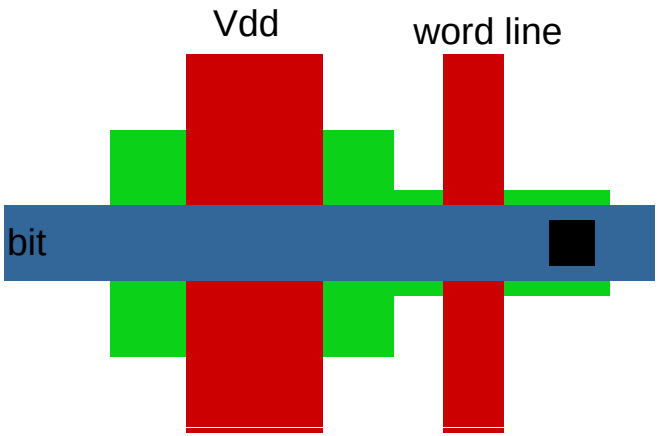
RAM – *DRAM Cells*

1-T DRAM cell



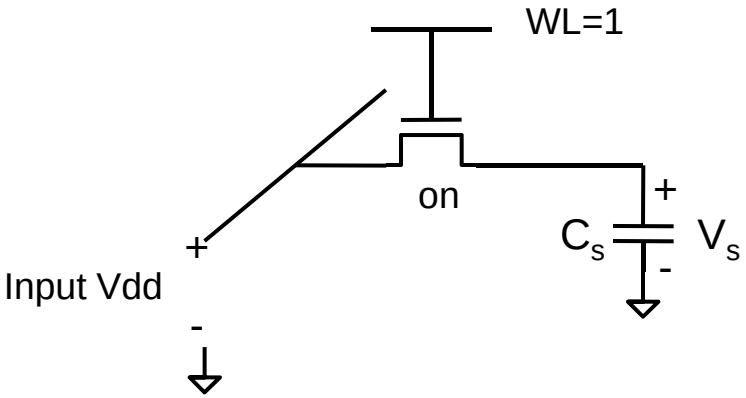


Layout of 1-T DRAM (right)

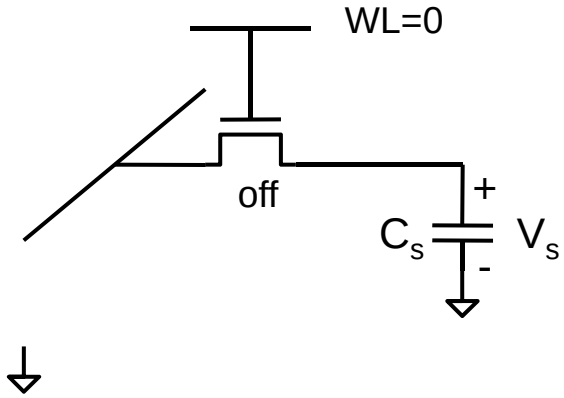


RAM – *DRAM Retention Time*

Write and hold operations in a DRAM cell



Write Operation



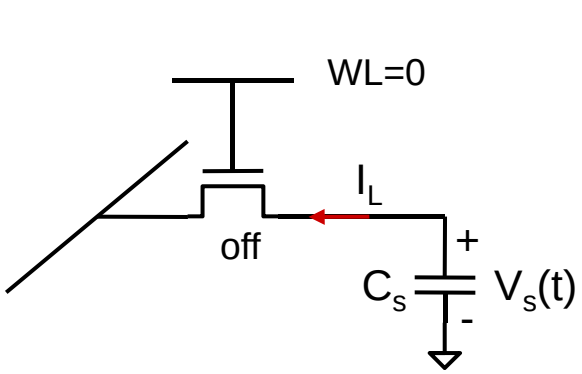
Hold

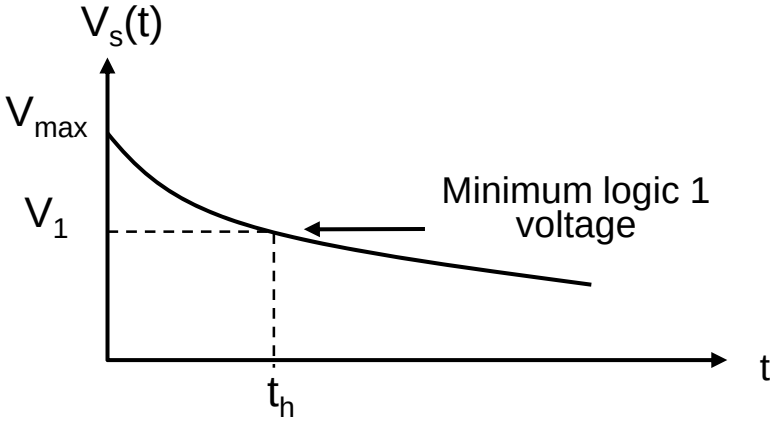
$$V_s = V_{\max} = V_{DD} - V_{tn}$$

$$Q_{\max} = C_s (V_{DD} - V_{tn})$$

RAM – *DRAM Retention Time*

Charge leakage in a DRAM Cell





$$I_L = -\left(\frac{dQ_s}{dt}\right)$$

$$I_L = -C_s\left(\frac{dV_s}{dt}\right)$$

$$I_L \approx -C_s\left(\frac{\Delta V_s}{\Delta t}\right)$$

$$t_h = |\Delta t| \approx \left(\frac{C_s}{I_L}\right)\Delta V_s$$

RAM – *DRAM Refresh Operation*

As an example, if IL=1nA, Cs=50fF, and the difference of Vs is 1V, the hold time is

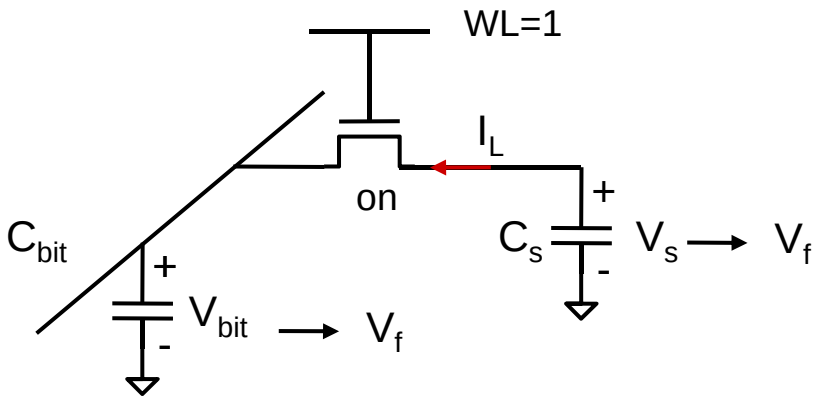
$$t_h = \frac{50 \times 10^{-15}}{1 \times 10^{-9}} \times 1 = 0.5 \mu s$$

Memory units must be able to hold data so long as the power is applied. To overcome the charge leakage problem, DRAM arrays employ a **refresh operation** where the data is periodically read from every cell, amplified, and rewritten.

The refresh cycle must be performed on every cell in the array with a minimum refresh frequency of about

$$f_{refresh} \approx \frac{1}{2t_h}$$

RAM – *DRAM Read Operation*



$$Q_s = C_s V_s$$

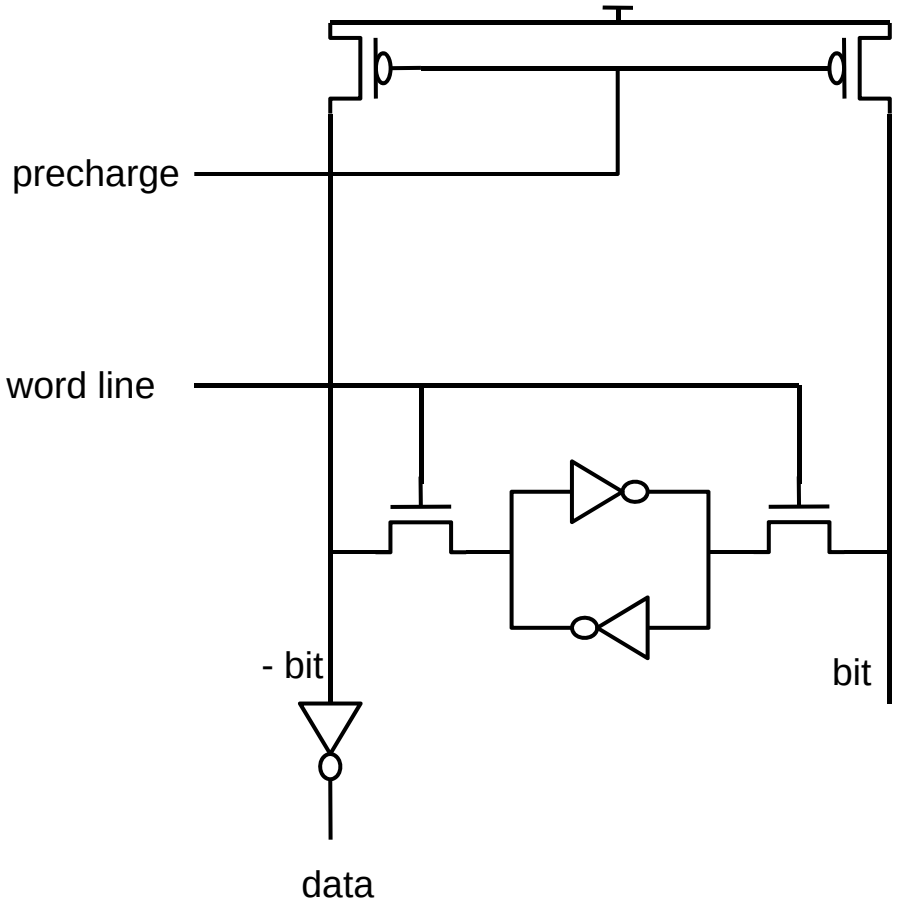
$$Q_s = C_s V_f + C_{bit} V_f$$

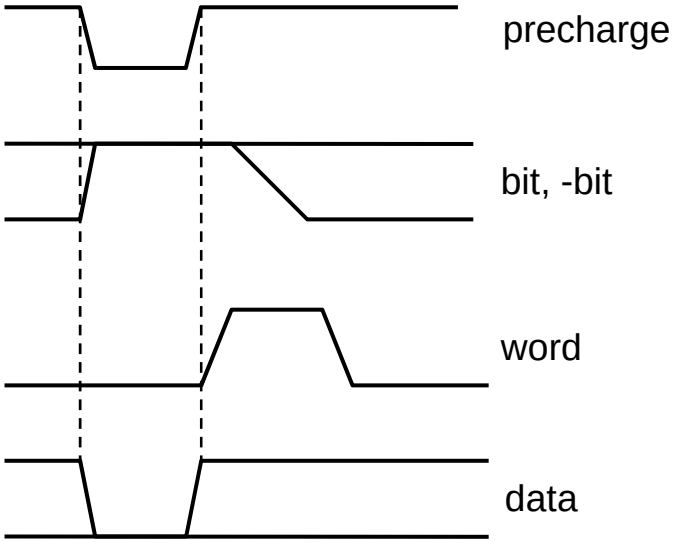
$$V_f = \left(\frac{C_s}{C_s + C_{bit}}\right) V_s$$

This shows that $V_f < V_s$ for a store logic 1. In practice, V_f is usually reduced to a few tenths of a volt, so that the design of the sense amplifier becomes a critical factor

RAM – *SRAM Read Operation*

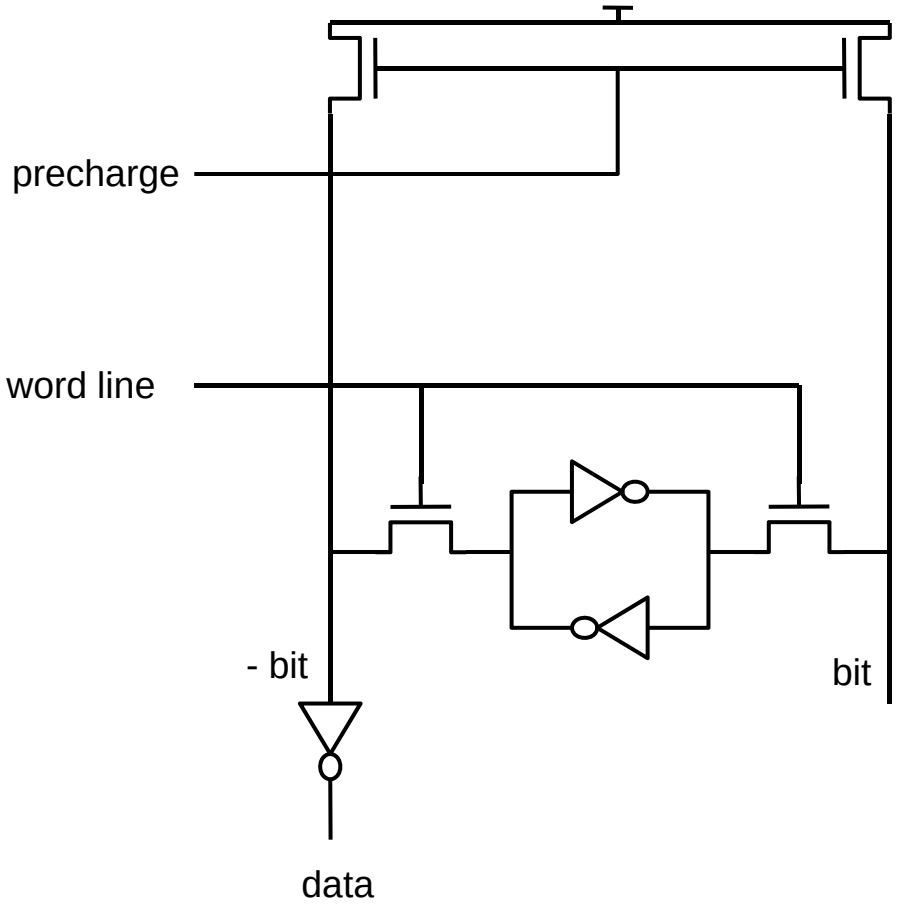
Vdd precharge

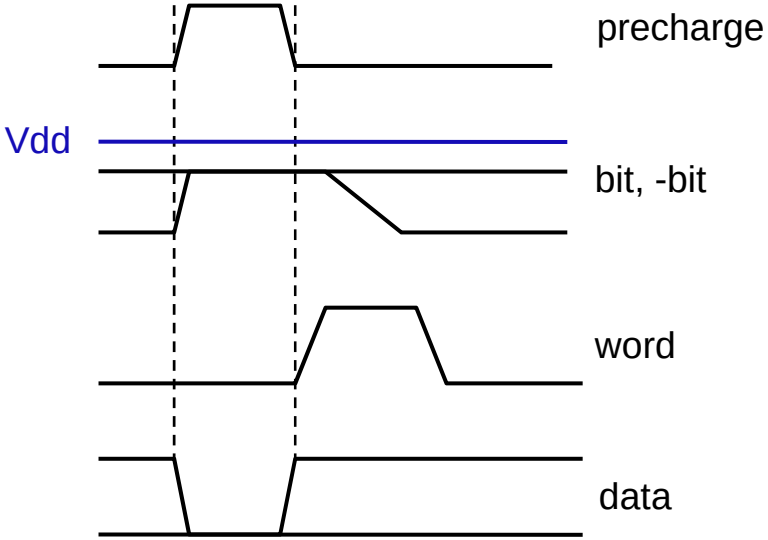




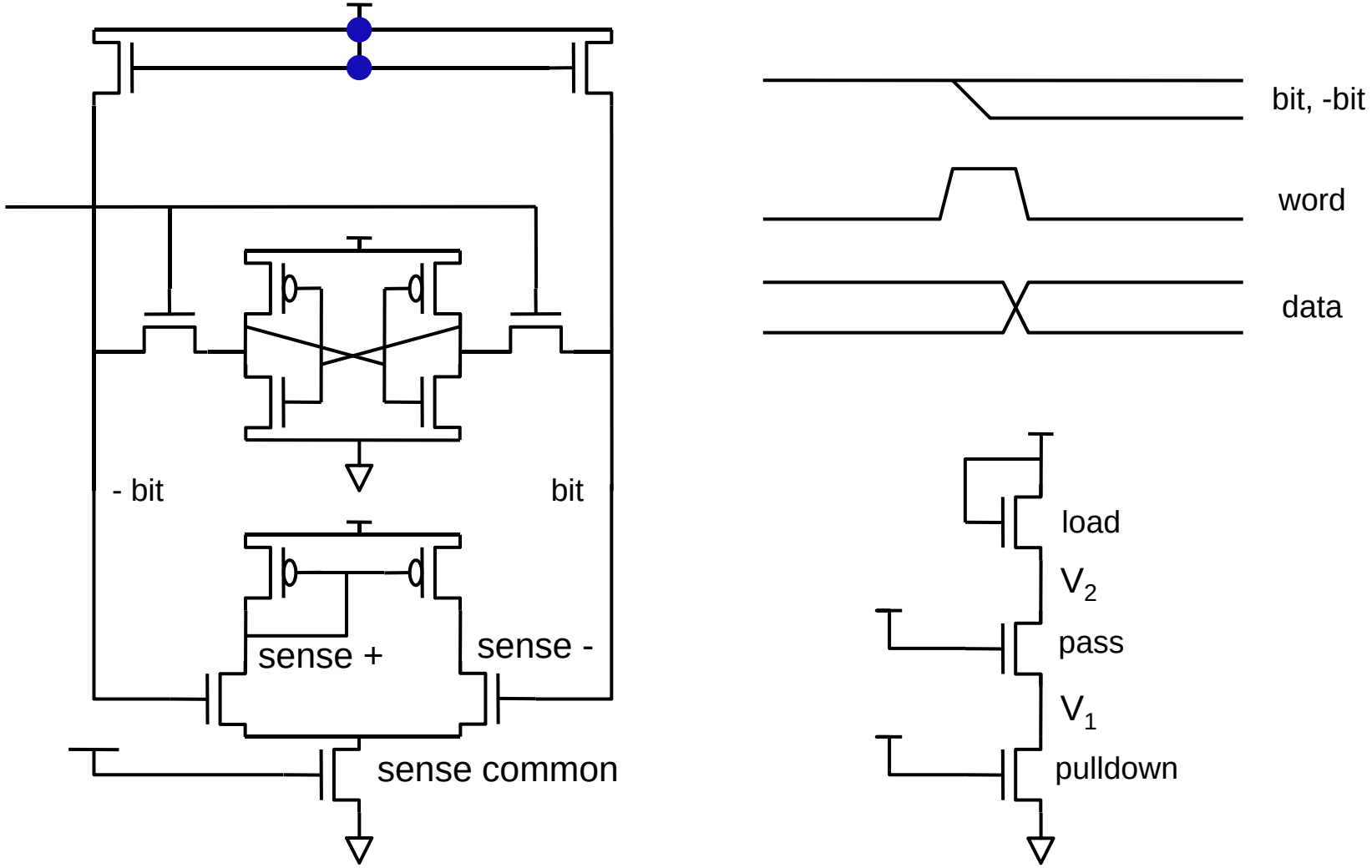
RAM – *SRAM Read Operation*

Vdd-Vtn precharge

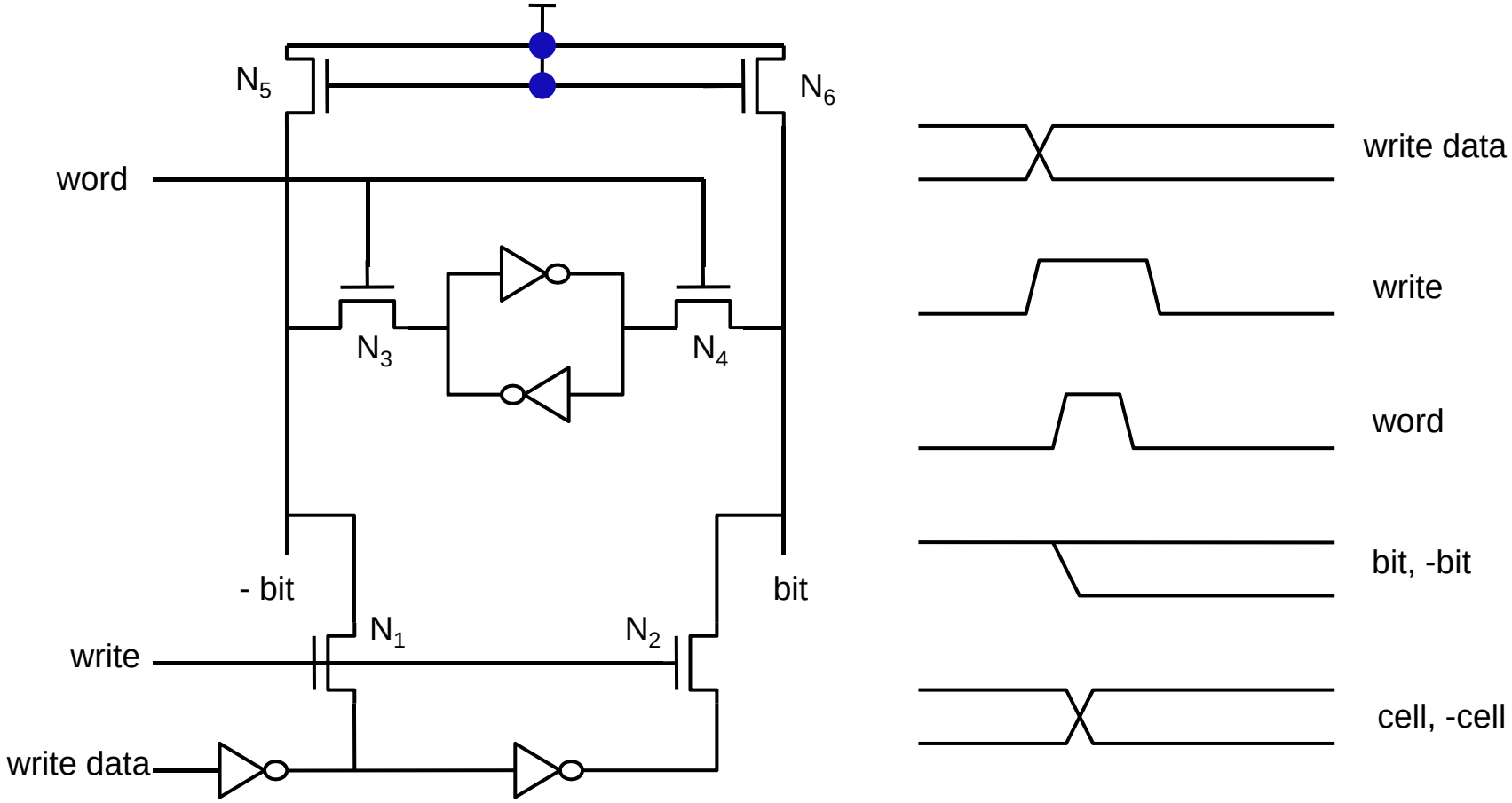




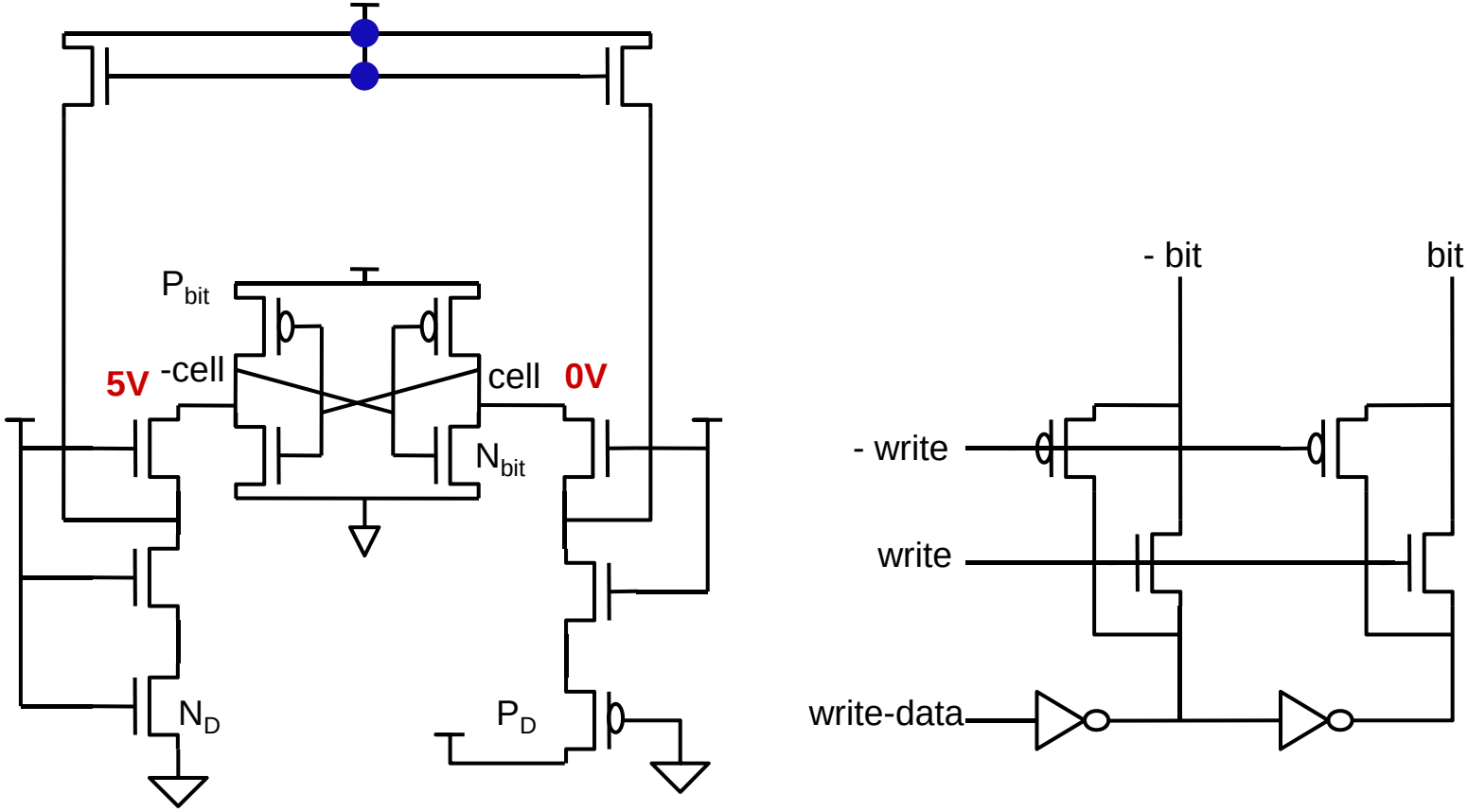
RAM – *SRAM Read Operation*



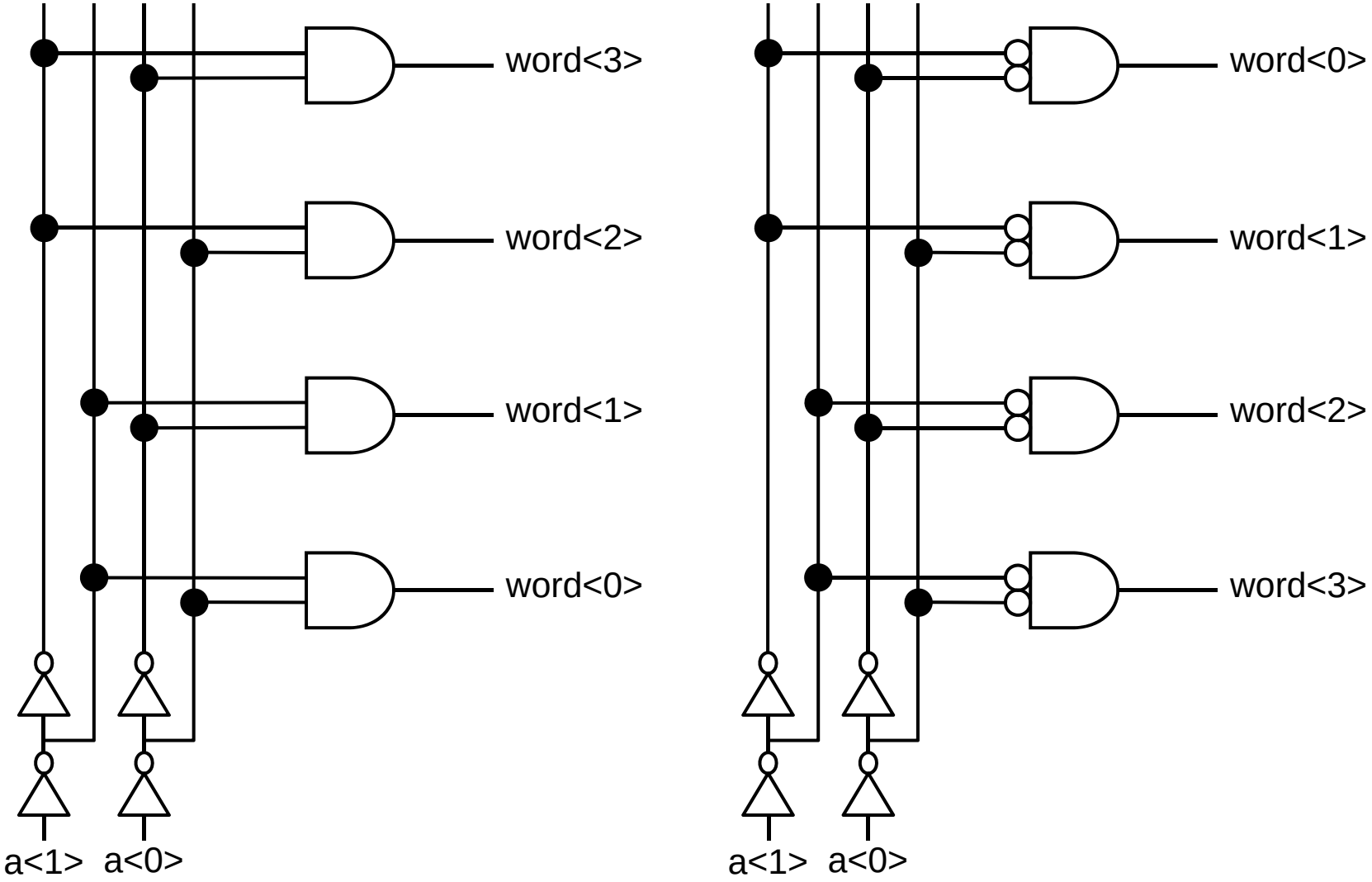
RAM – *Write Operation*



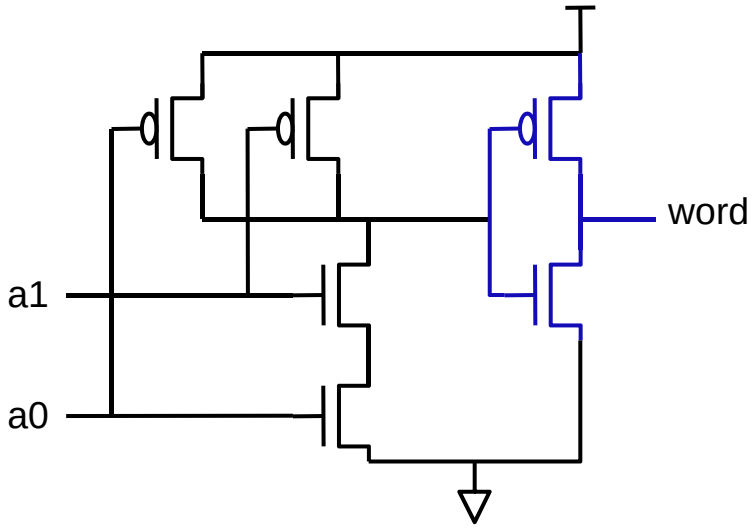
RAM – *Write Operation*



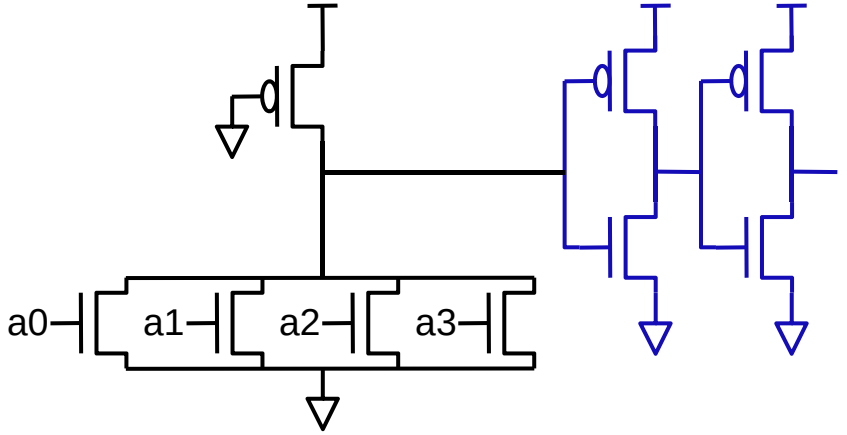
RAM – *Row Decoder*



RAM – *Row Decoder*



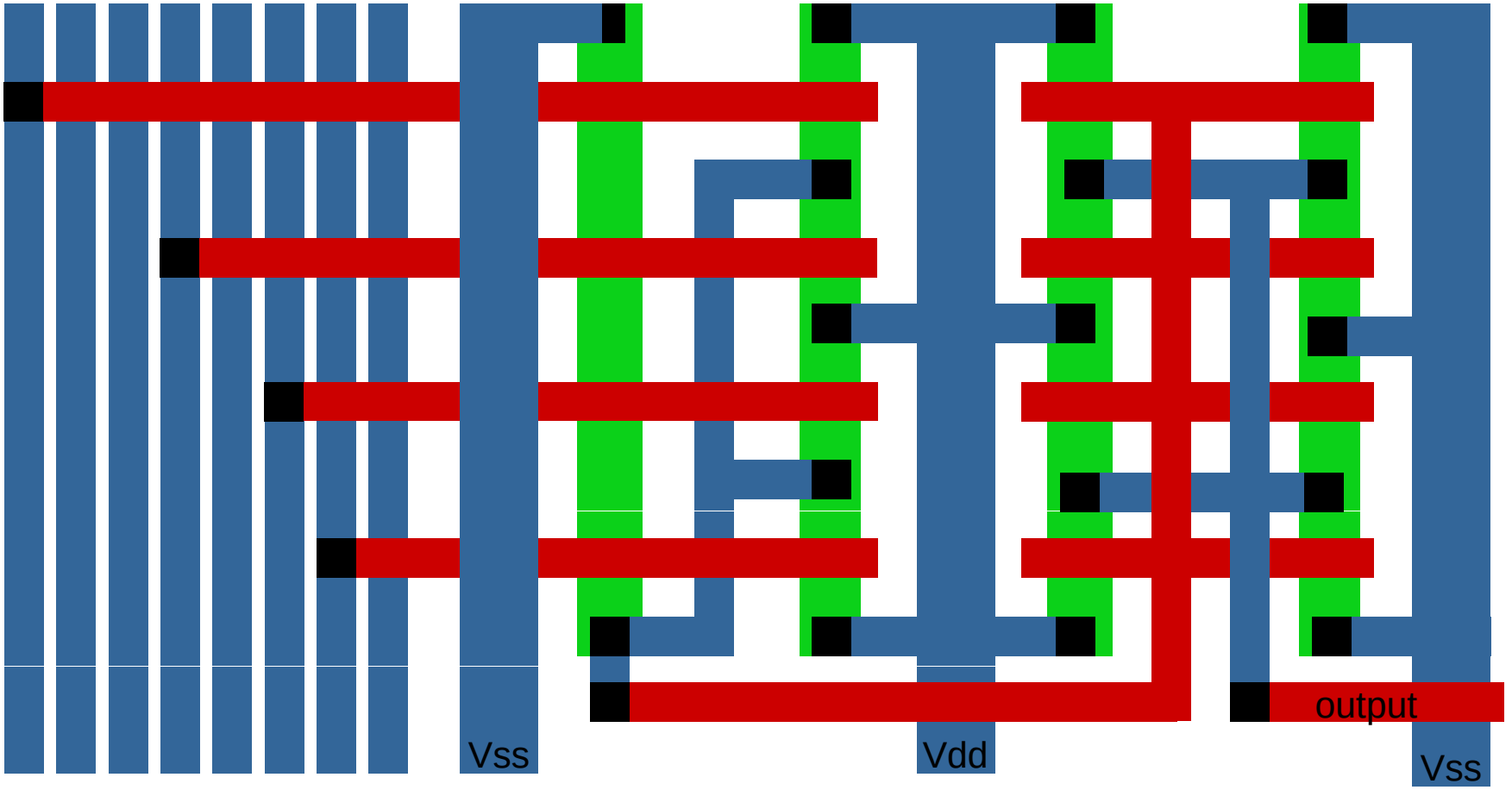
Complementary AND gate



Pseudo-nMOS gate

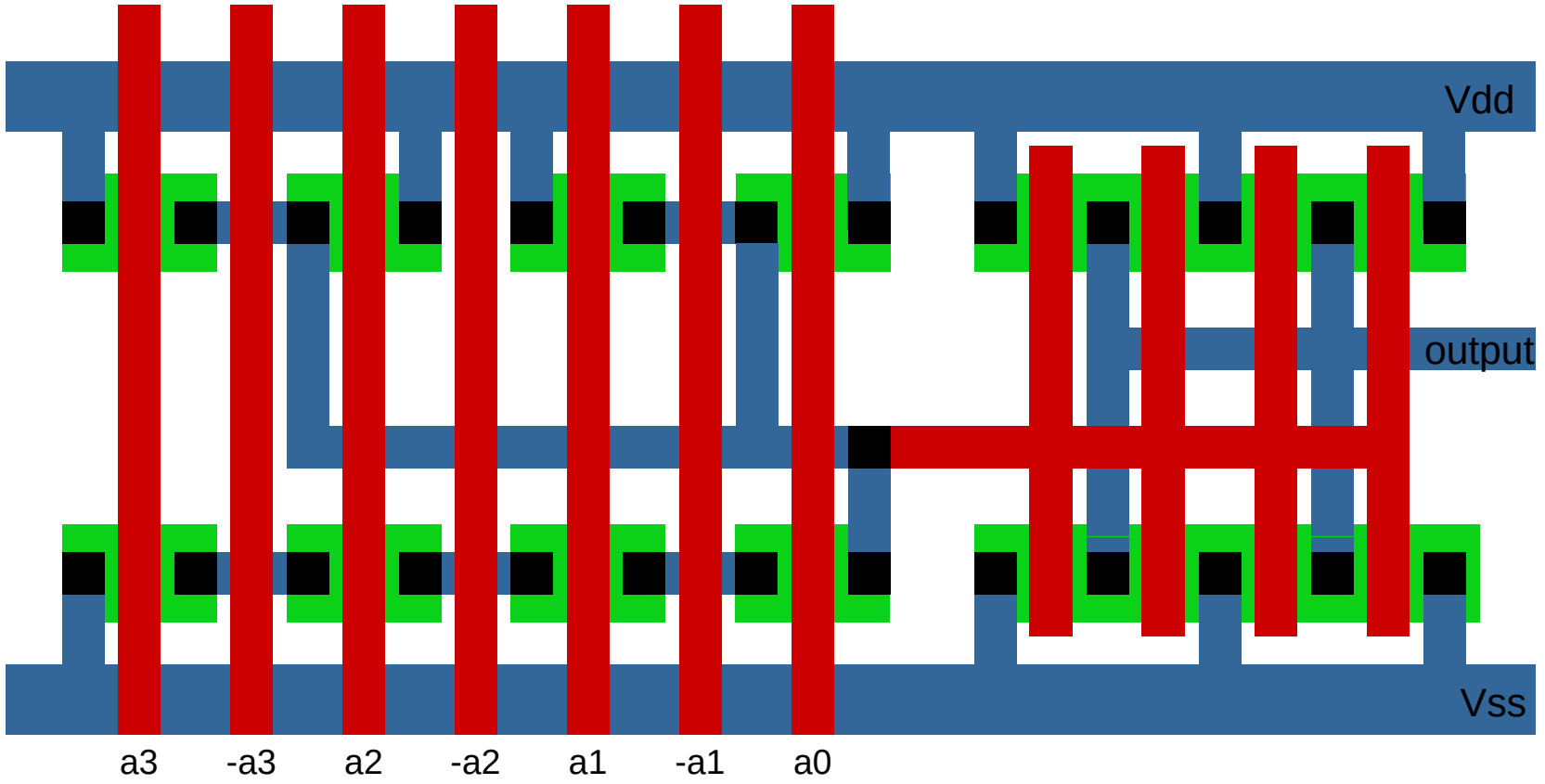
RAM – *Row Decoder*

Symbolic layout of row decoder



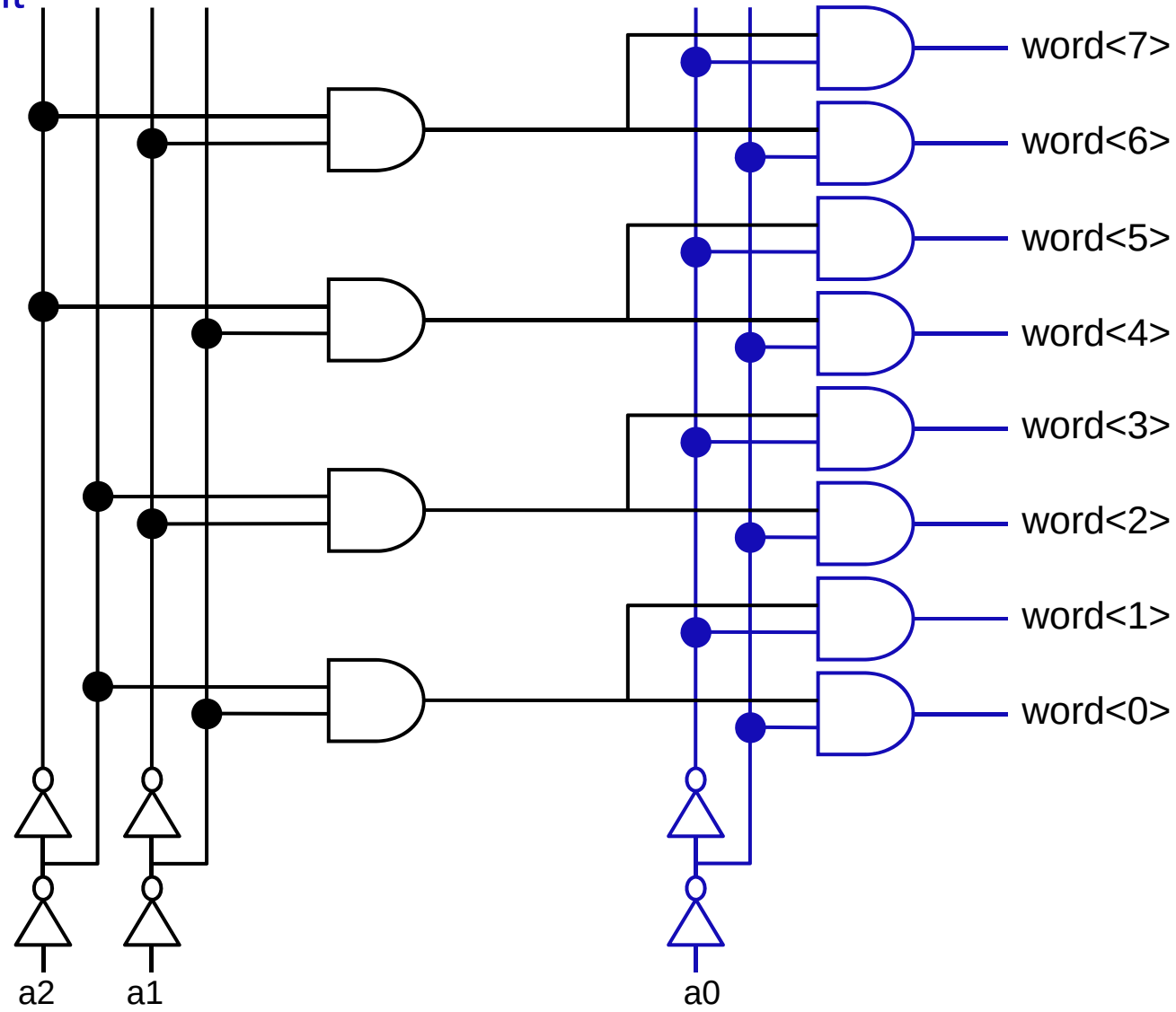
RAM – *Row Decoder*

Symbolic layout of row decoder



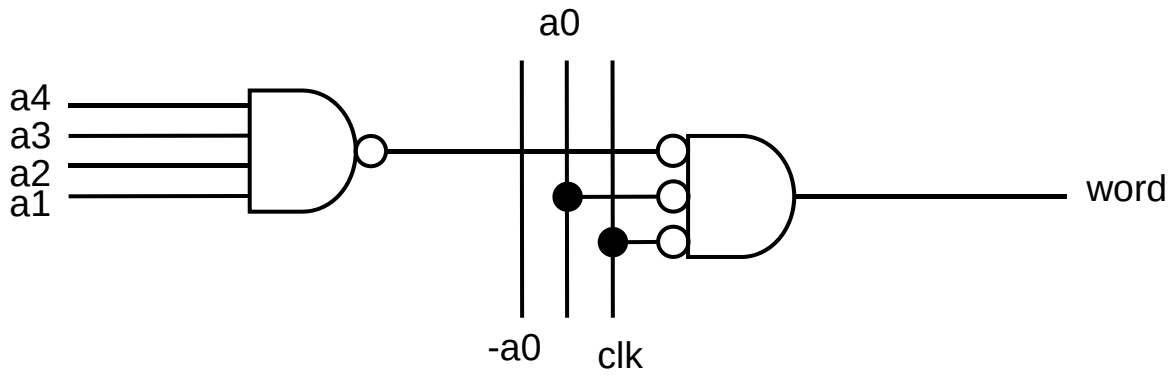
RAM – *Row Decoder*

Predecode circuit

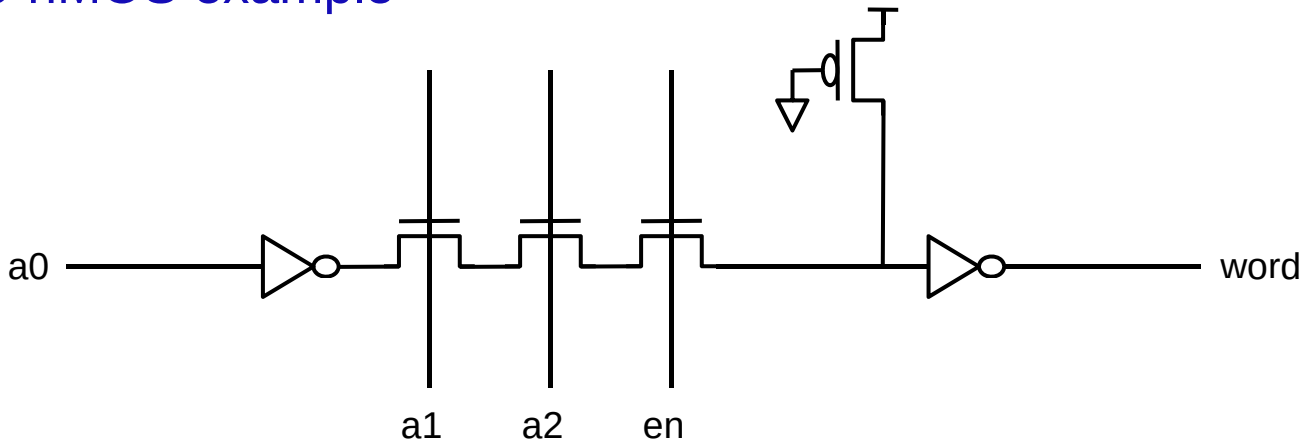


RAM – *Row Decoder*

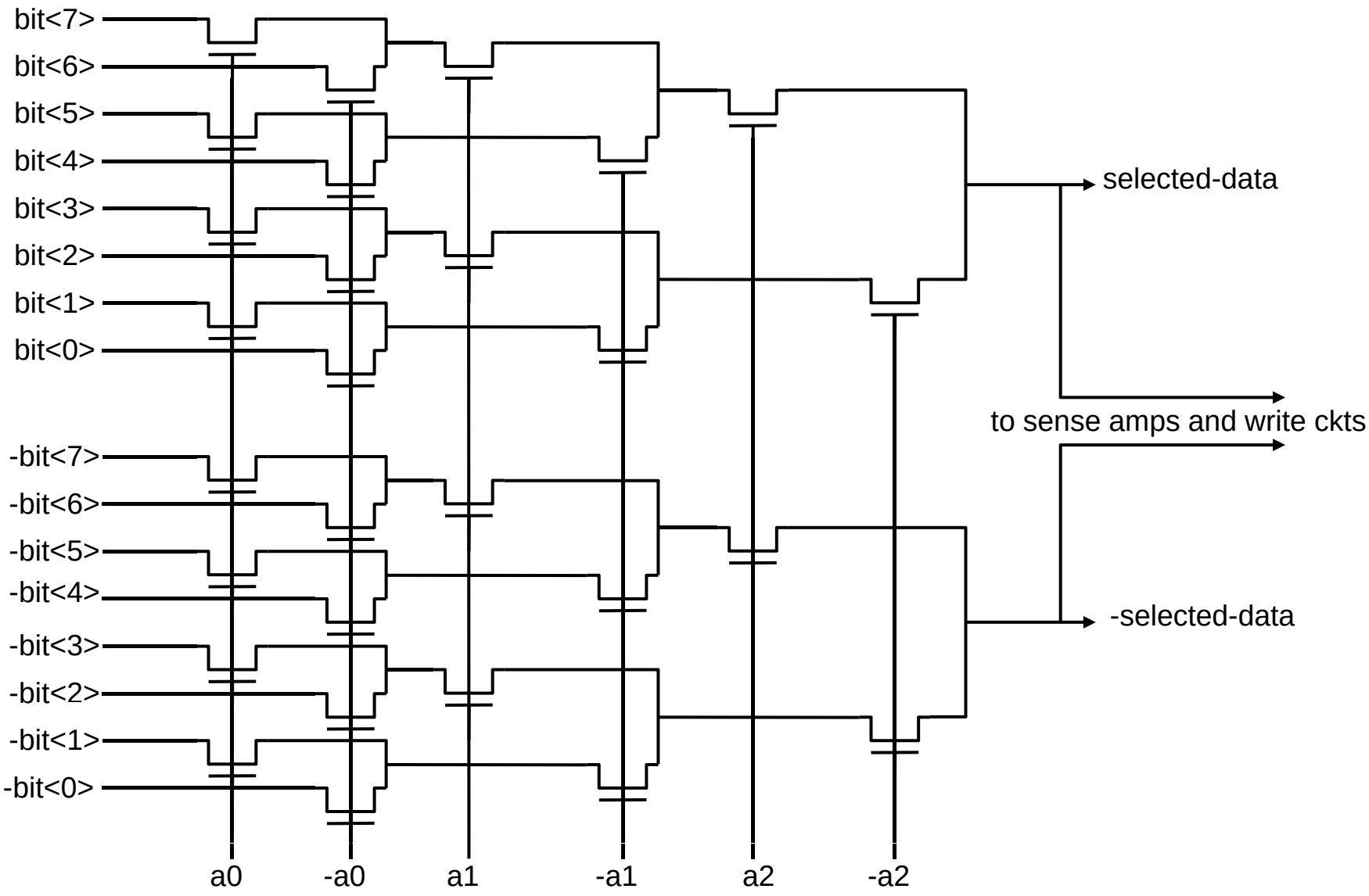
Actual implementation



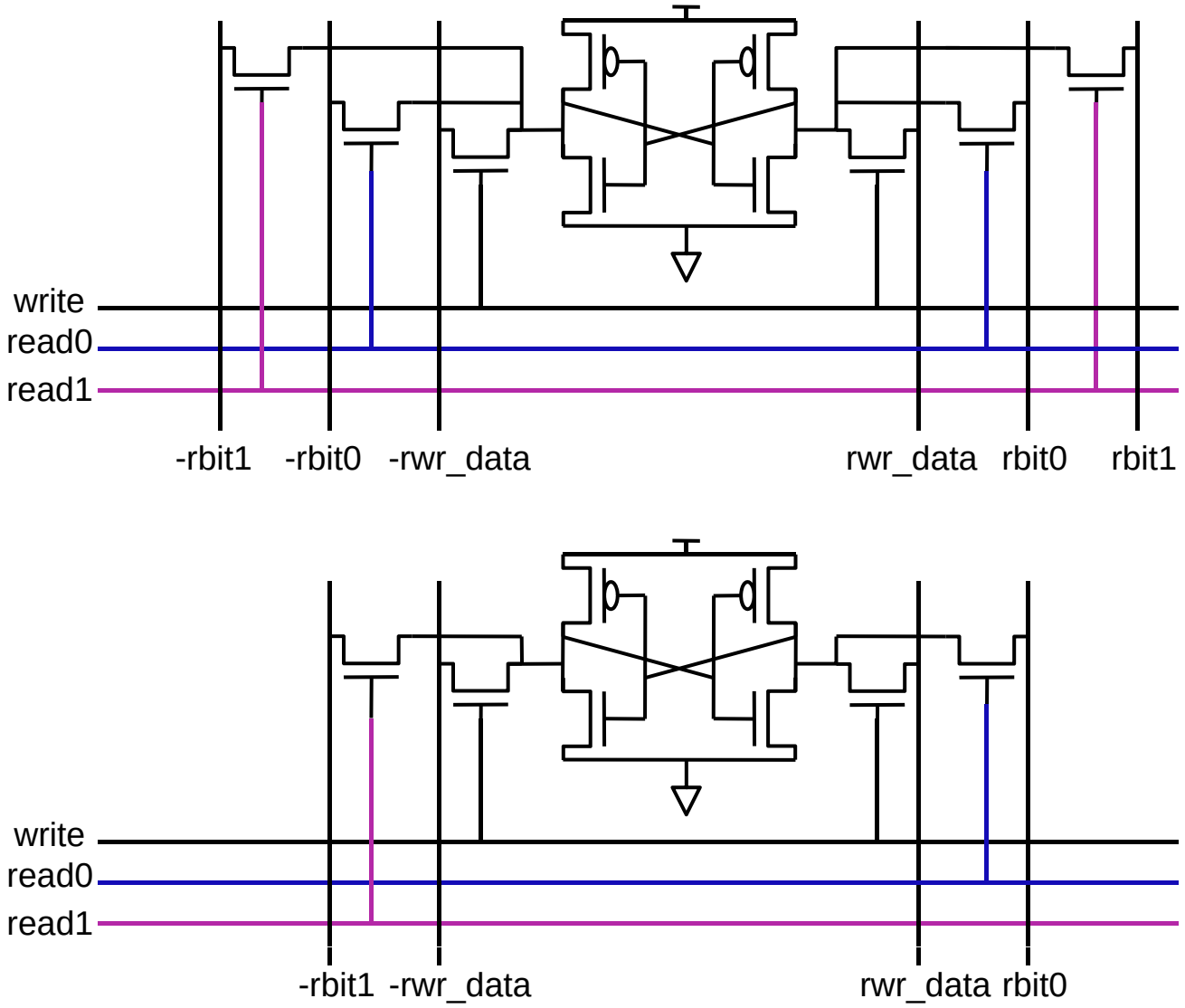
Pseudo-nMOS example



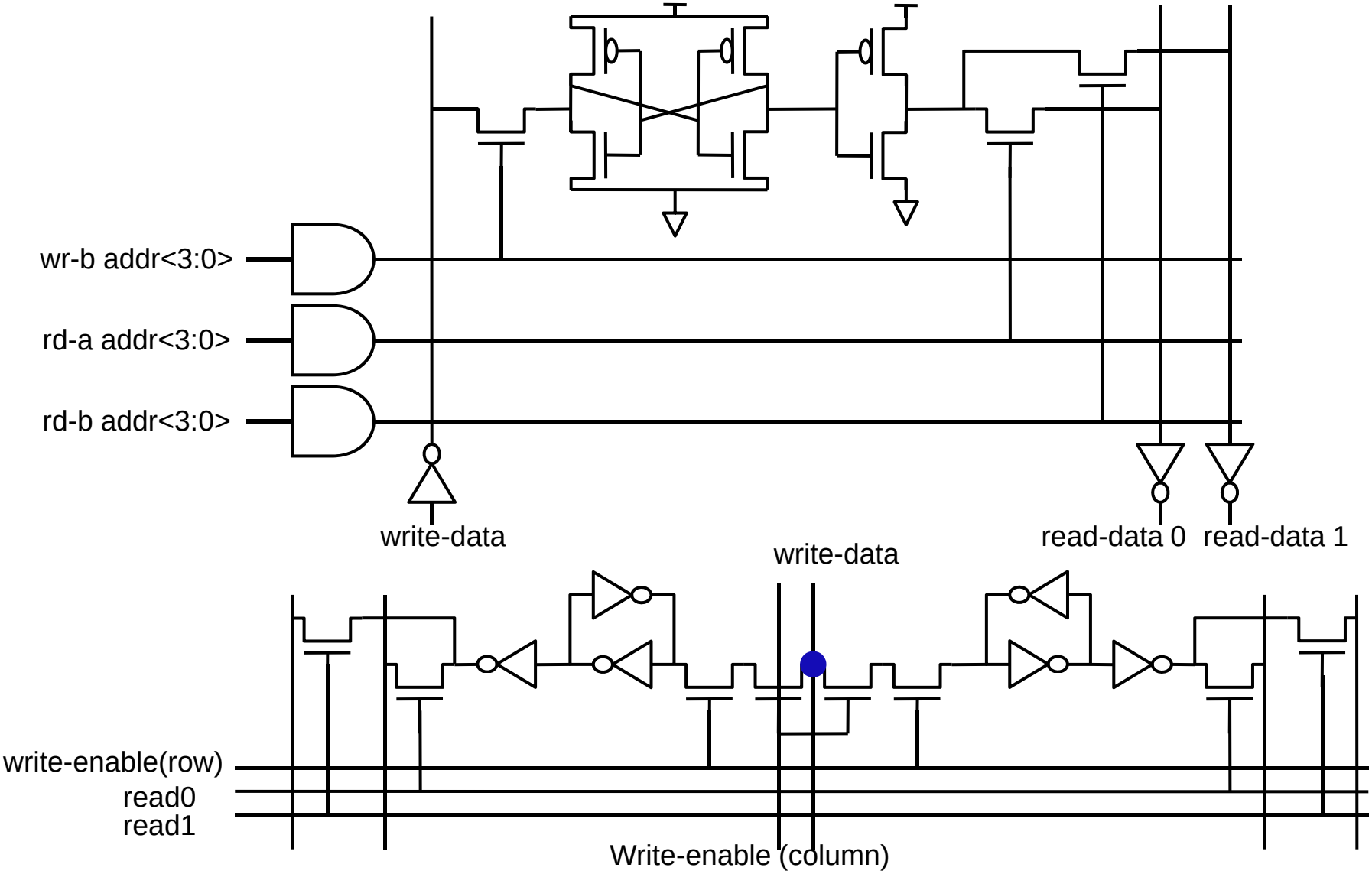
RAM – *Column Decoder*



RAM – *Multi-port RAM*

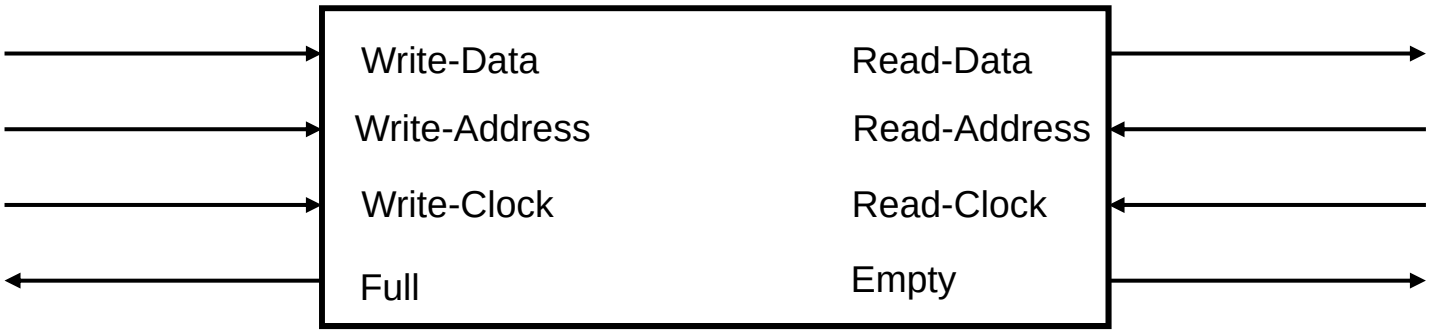


RAM – *Expandable Reg. File Cell*

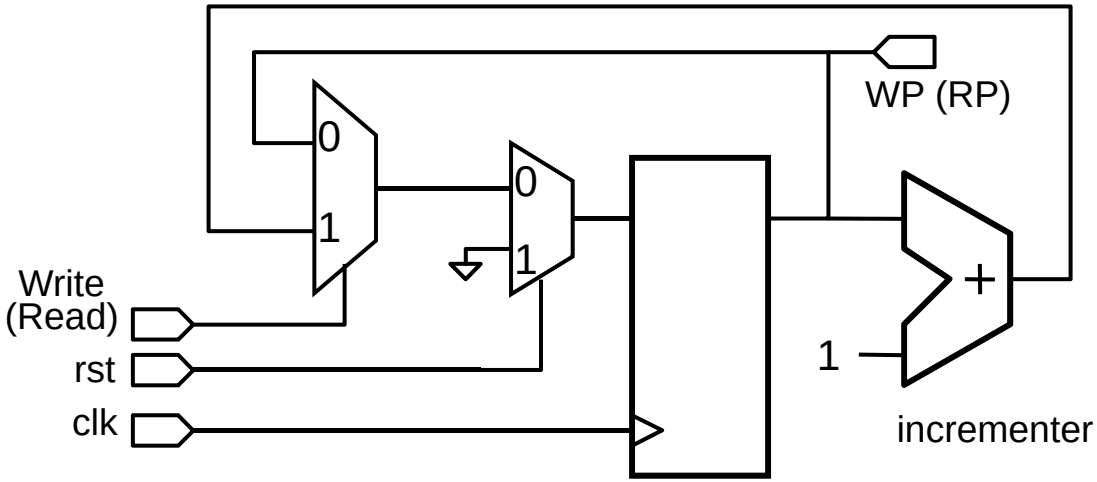


Specific Memory – *FIFO*

Two port RAM

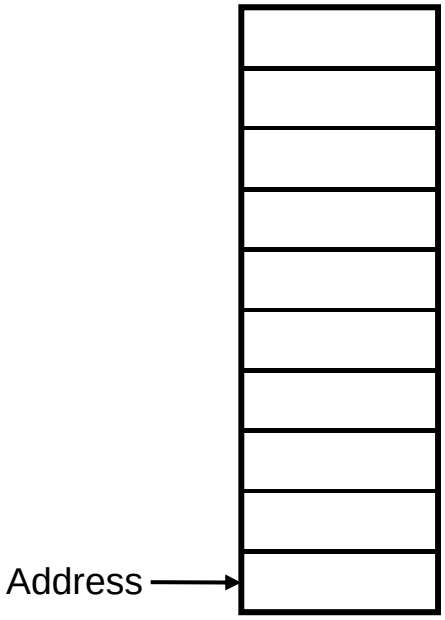


FIFO Write (Read) address control design



Specific Memory – *LIFO*

LIFO (Stack)



Require:

Single port RAM

One address counter

Empty/Full detector

Algorithm:

Write: write current address

Address=Address+1

Read: Address=Address-1

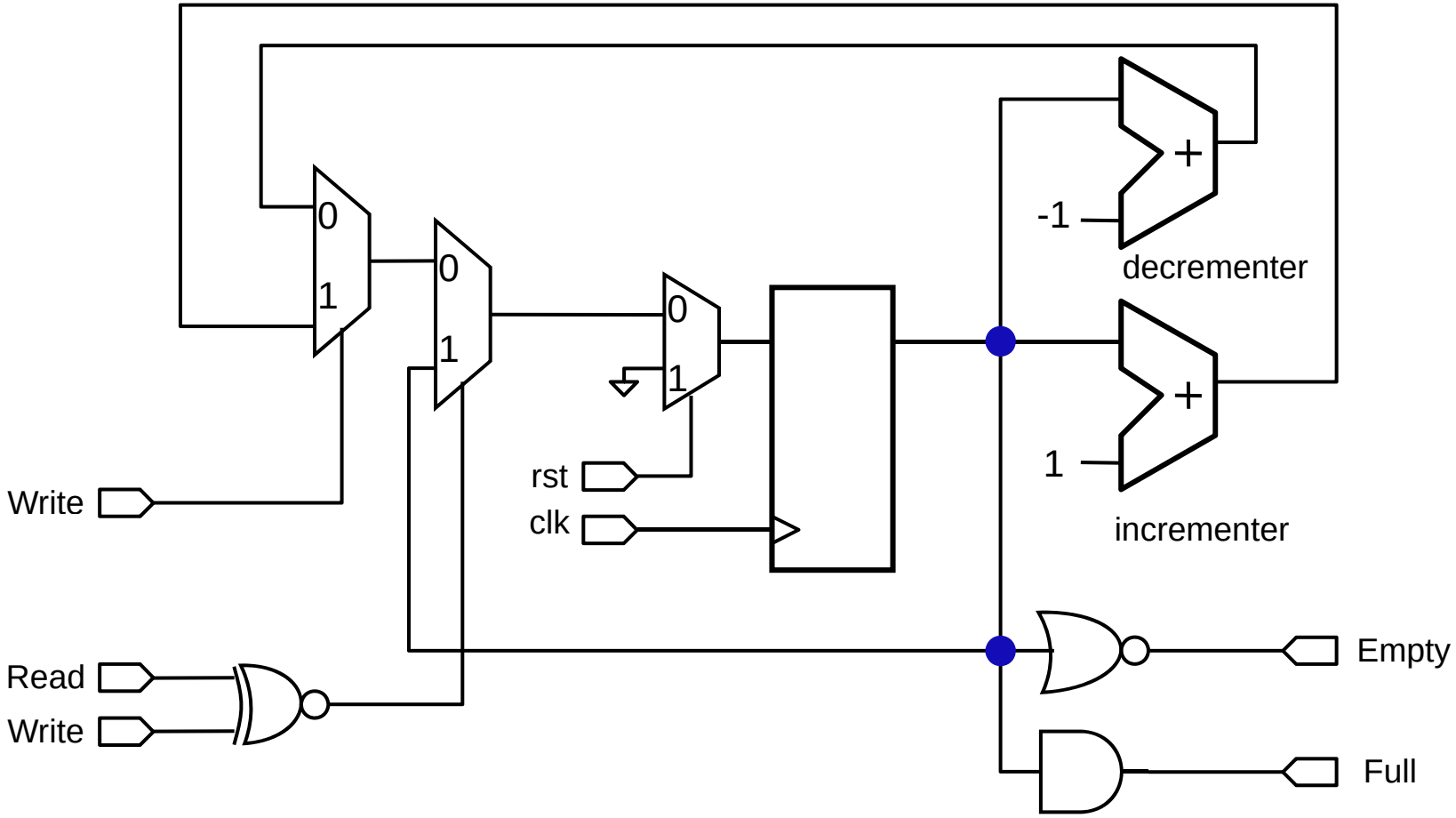
read current address

Empty: Address=0

Full : Address=FFF…

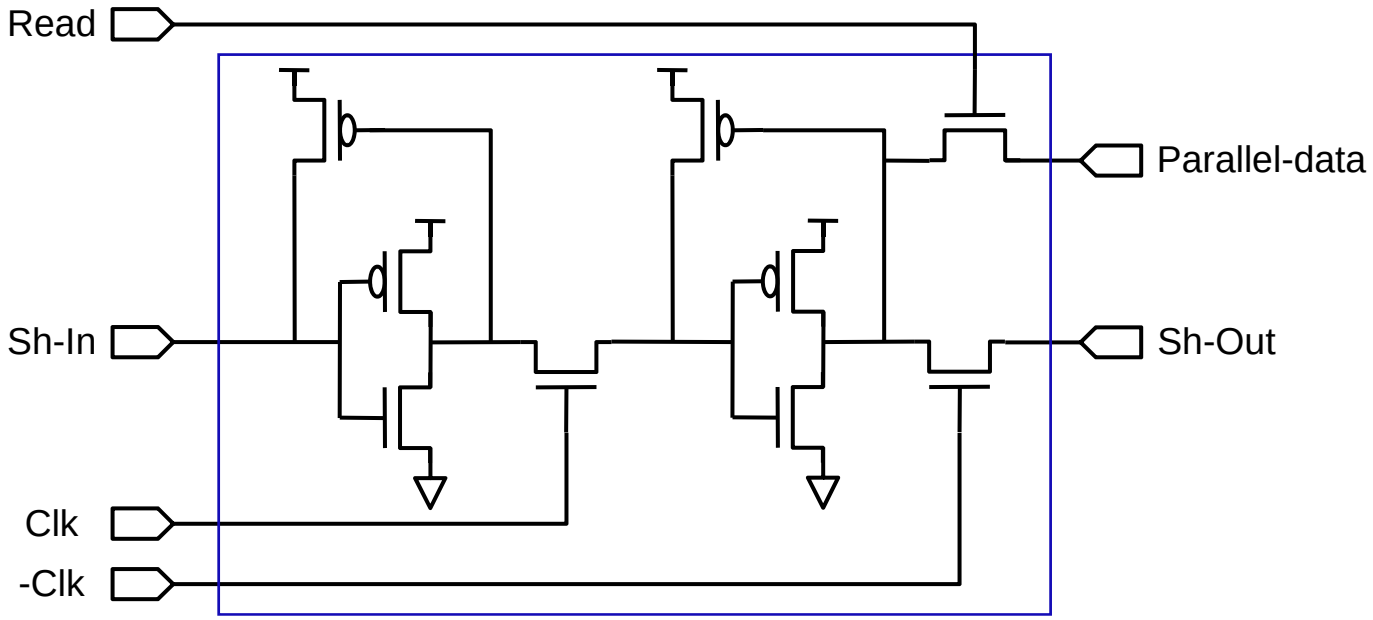
Specific Memory – *LIFO*

LIFO address control design

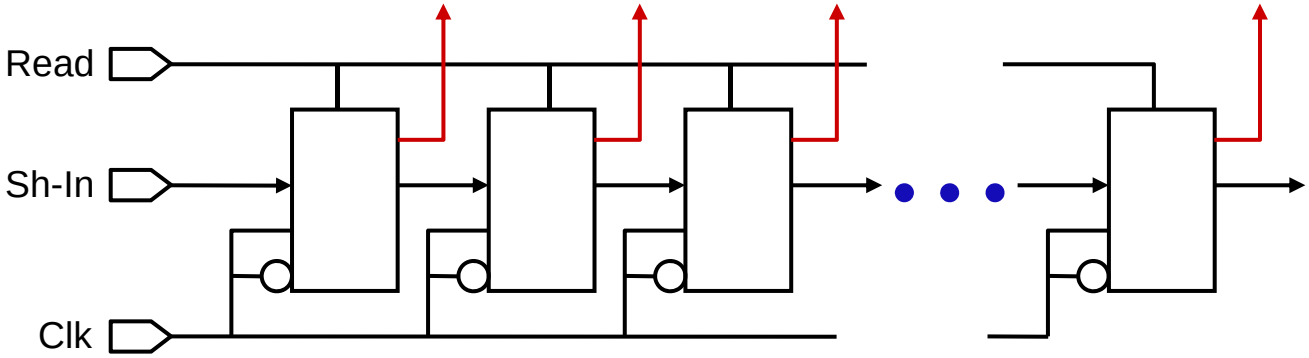


Specific Memory – *SIPO*

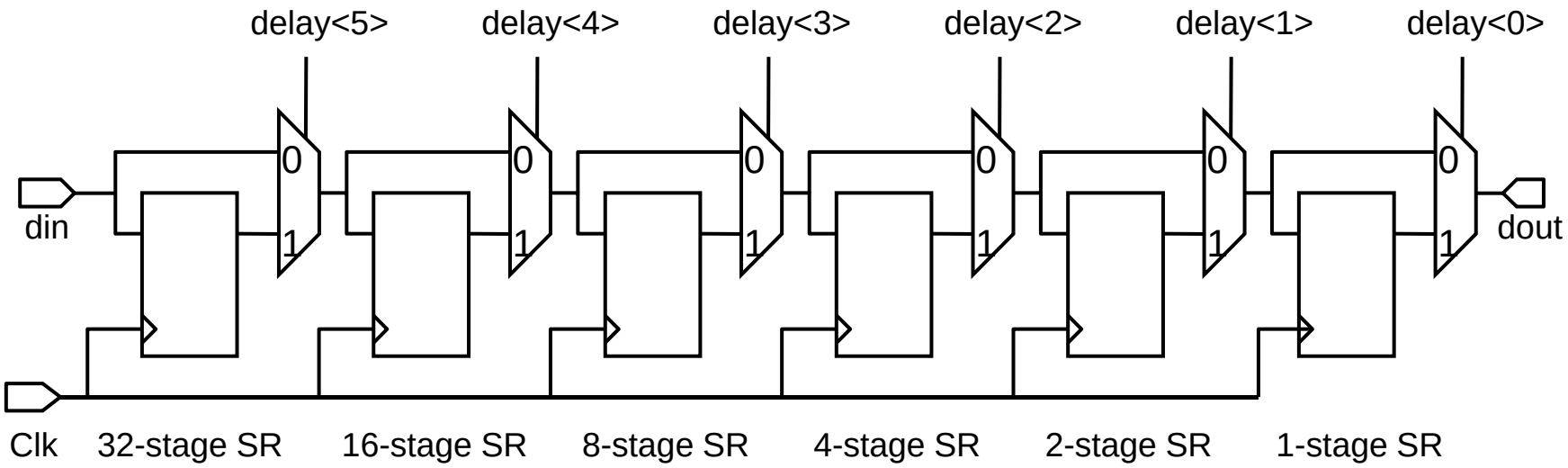
SIPO cell design



SIPO



Specific Memory – *Tapped Delay Line*

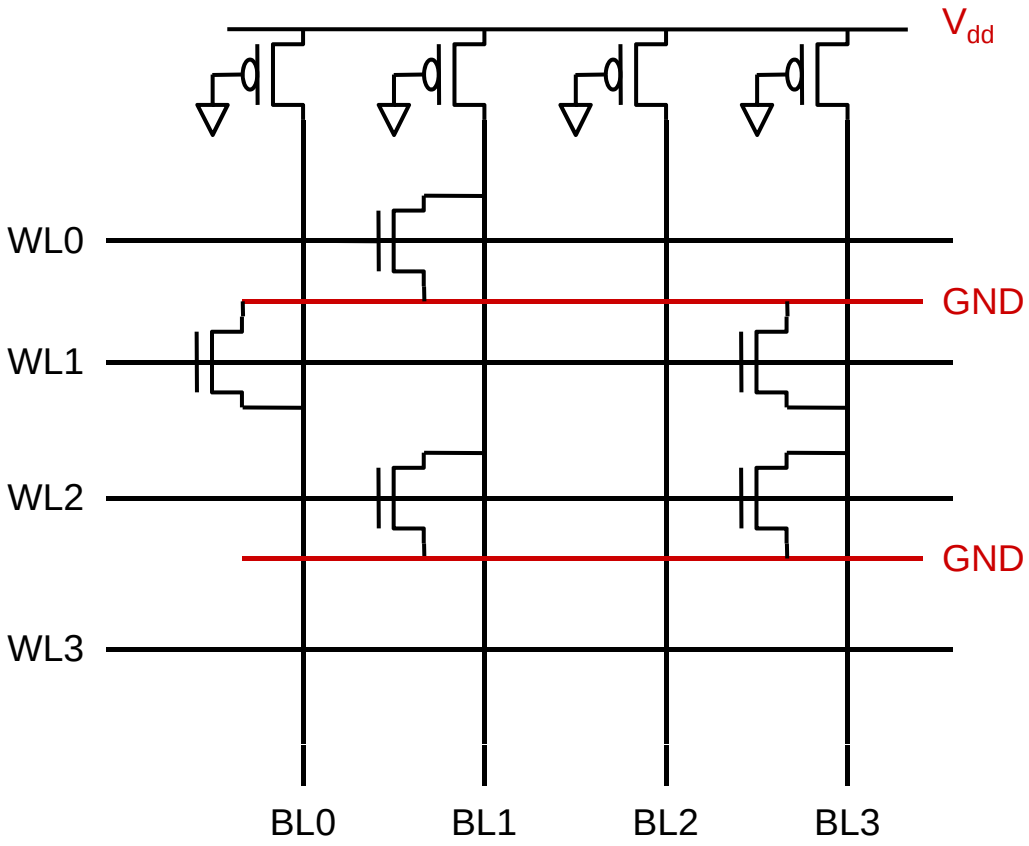


Read-Only Memories

- Read-Only Memories are nonvolatile
  - Retain their contents when power is removed
- Mask-programmed ROMs use one transistor per bit
  - Presence or absence determines 1 or 0

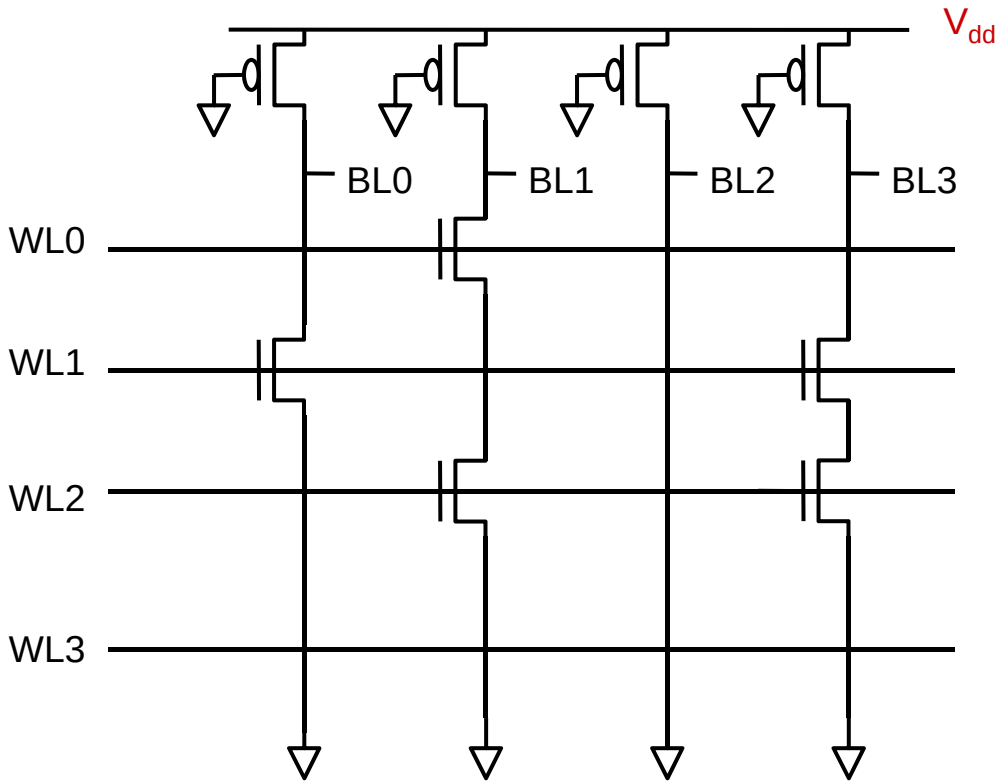
Non-volatile Memory – *ROM*

4x4 NOR-type ROM



Non-volatile Memory – *ROM*

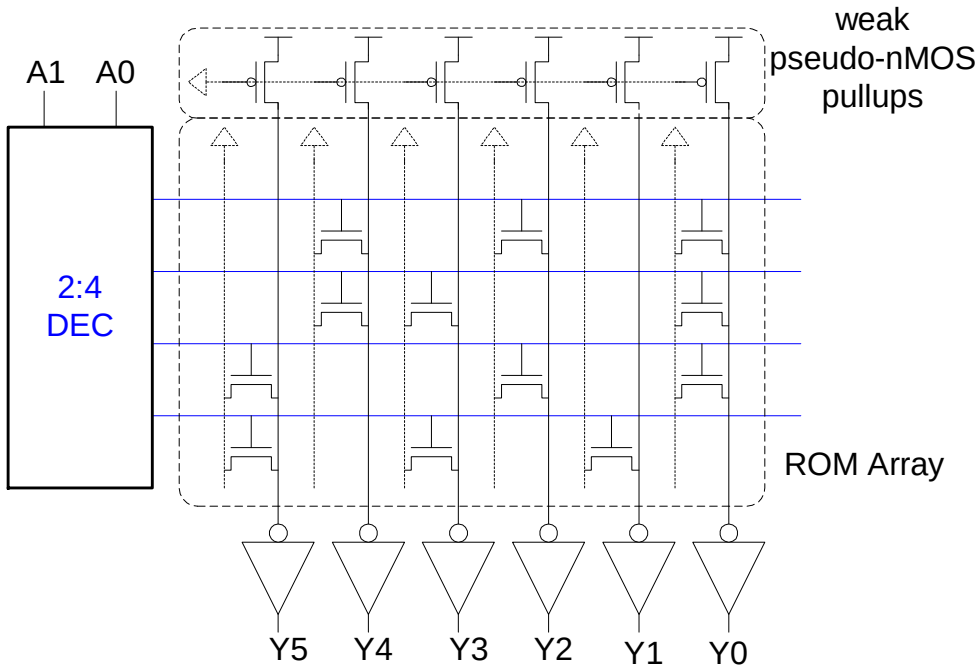
4x4 NAND-type ROM

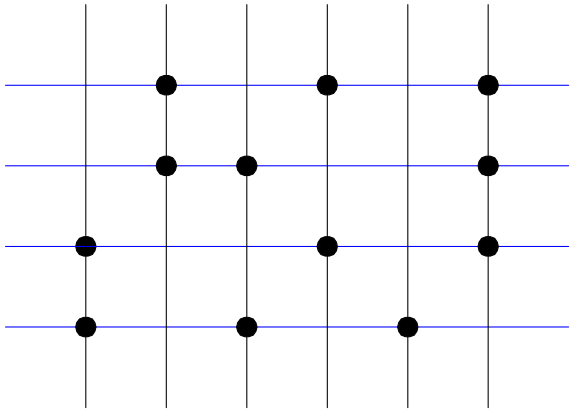


ROM Example

- 4-word x 6-bit ROM
  - Represented with dot diagram
  - Dots indicate 1's in ROM

```
Word 0: 010101
Word 1: 011001
Word 2: 100101
Word 3: 101010
```

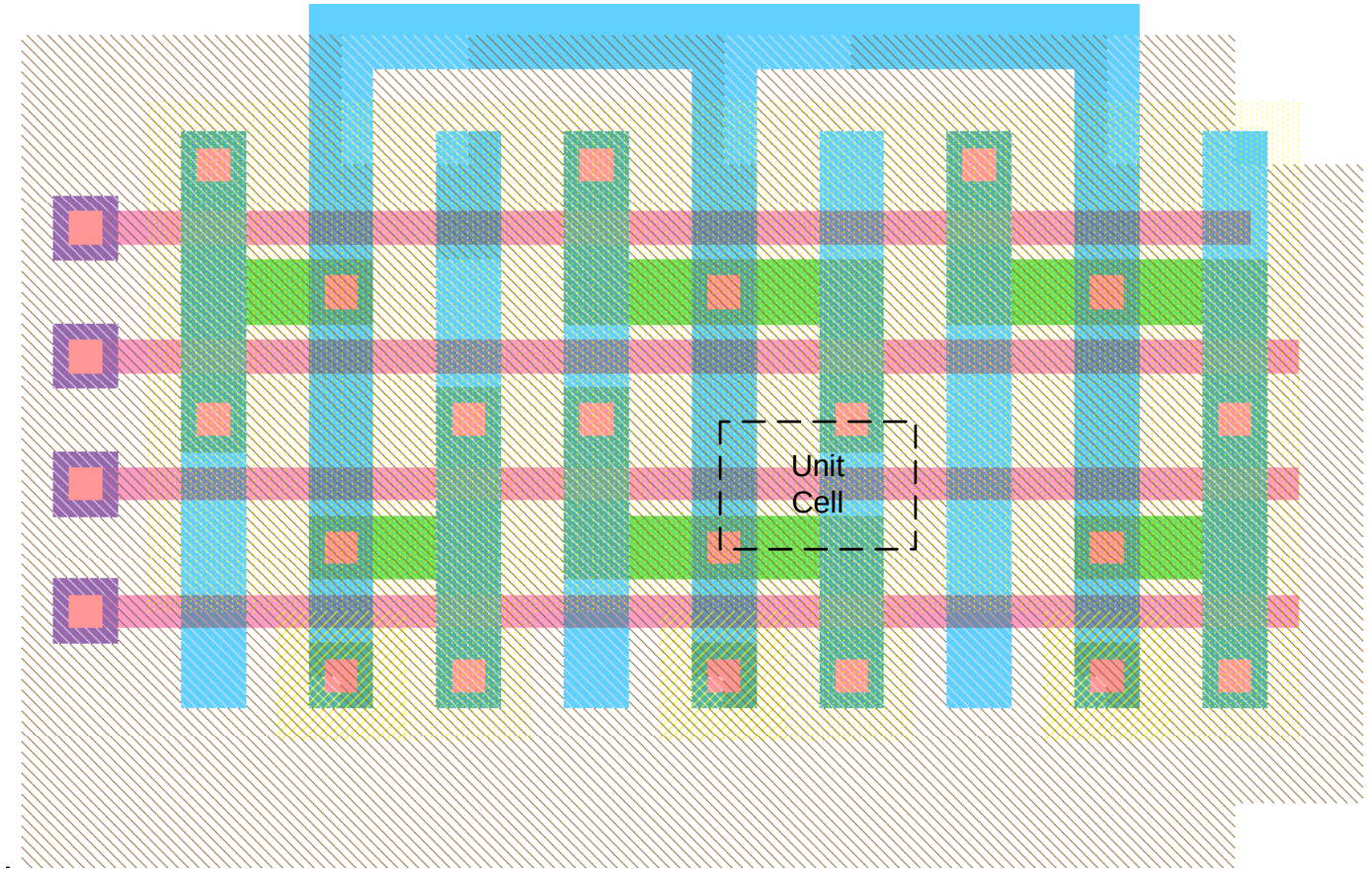




Looks like 6 4-input pseudo-nMOS NORs

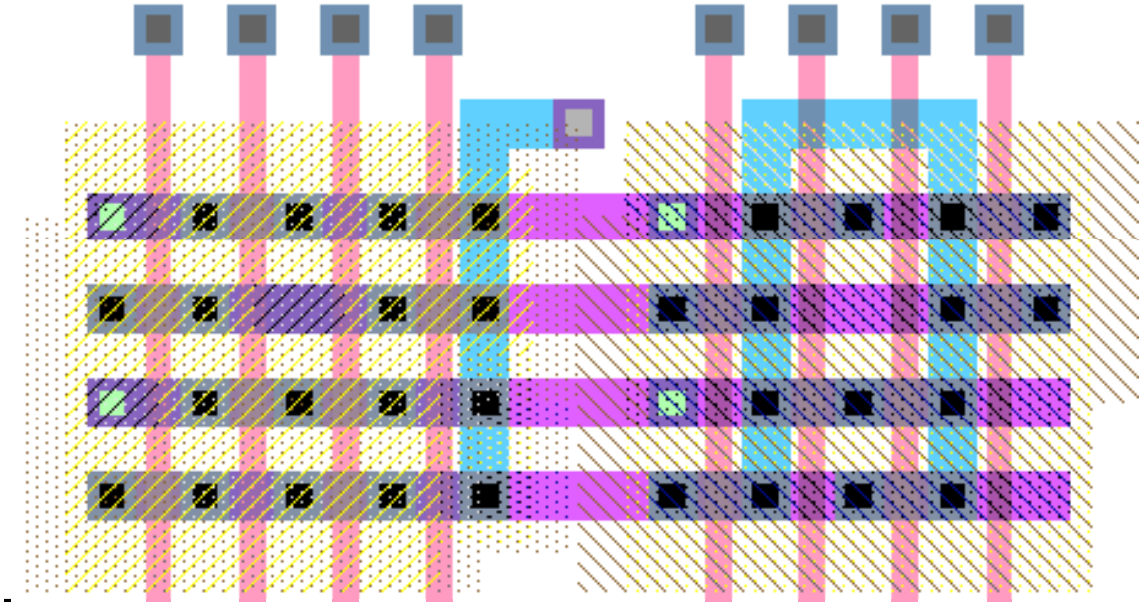
ROM Array Layout

- Unit cell is 12 x 8 λ (about 1/10 size of SRAM)

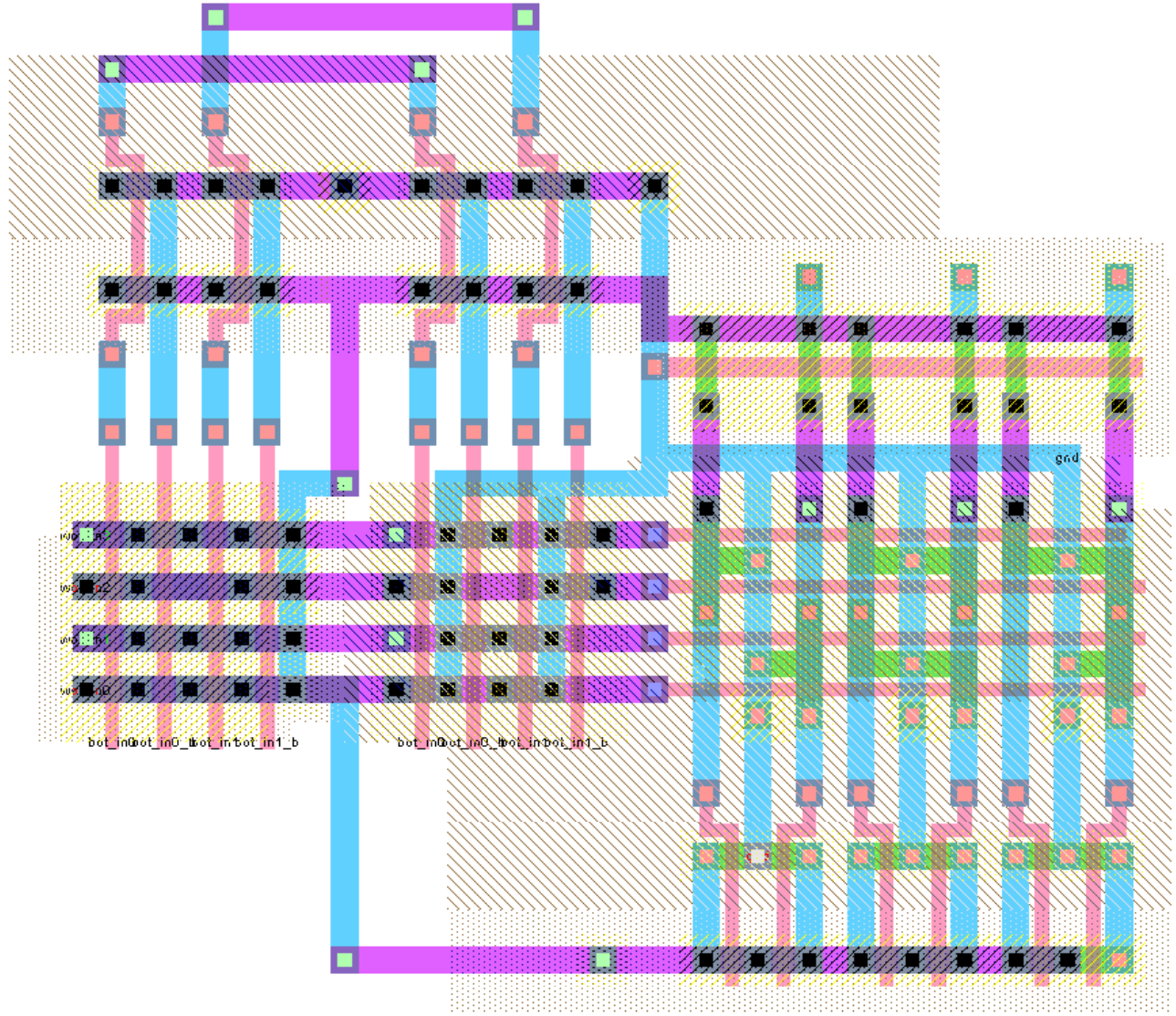


Row Decoders

- ROM row decoders must pitch-match with ROM
  - Only a single track per word!

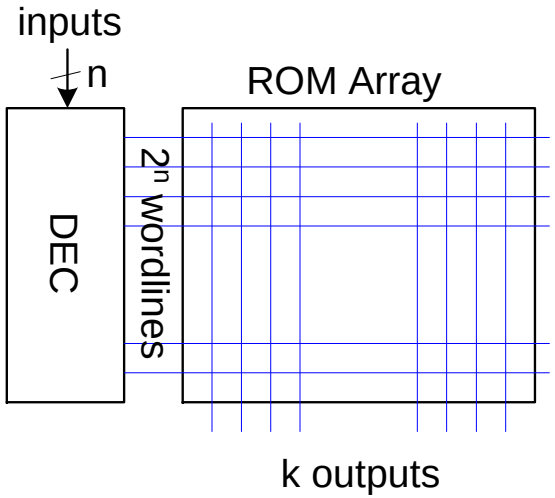


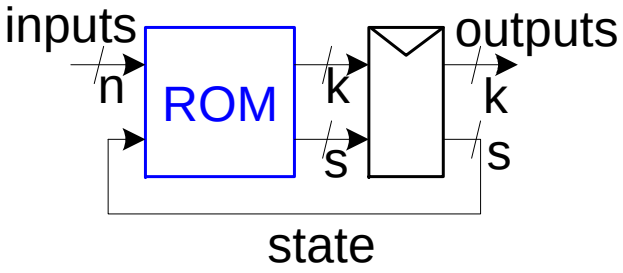
Complete ROM Layout



Building Logic with ROMs

- Use ROM as lookup table containing truth table
  - n inputs, k outputs requires 2^n words x k bits
  - Changing function is easy – reprogram ROM
- Finite State Machine
  - n inputs, k outputs, s bits of state
  - Build with 2^{n+s} x (k+s) bit ROM and (k+s) bit reg

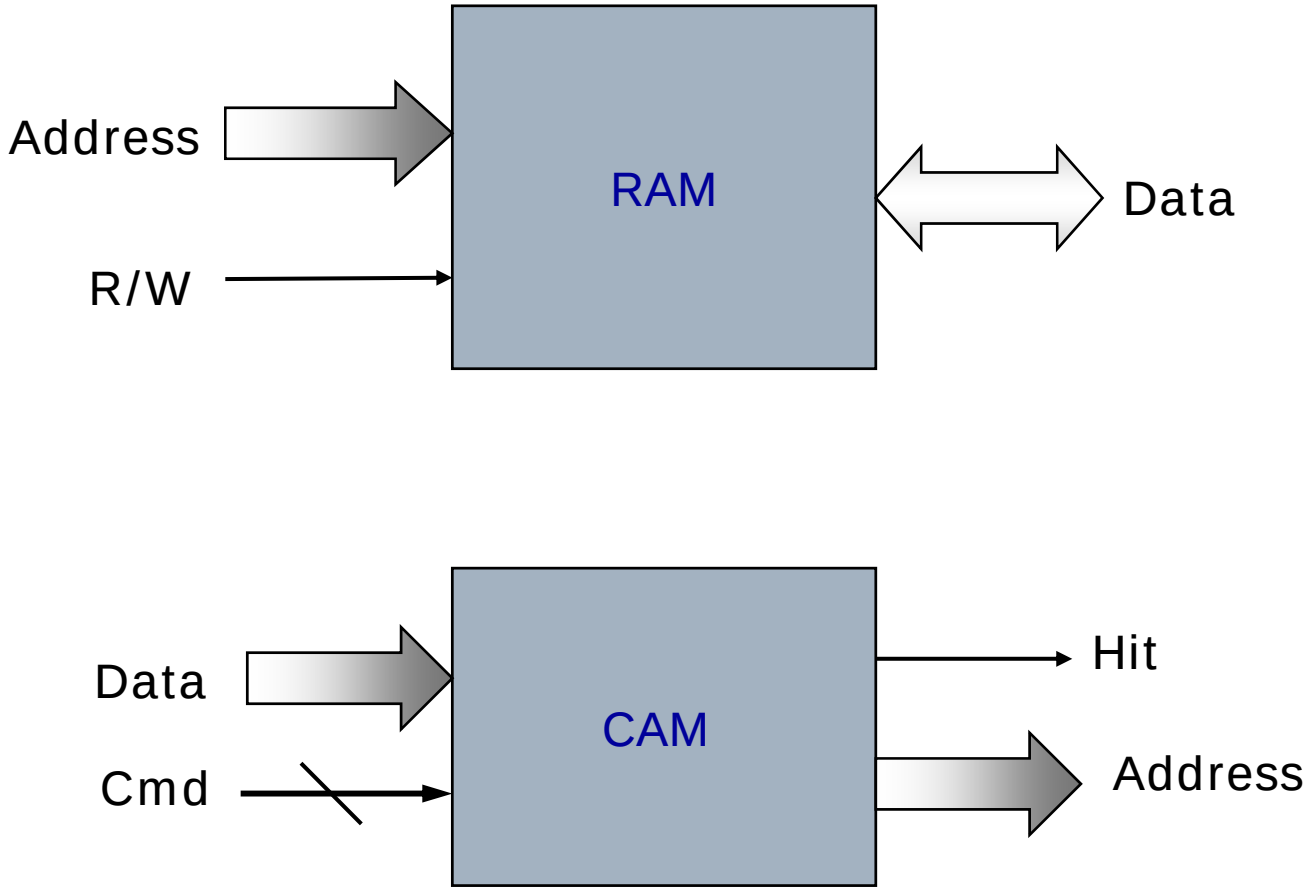




Specific Memory – *CAM*

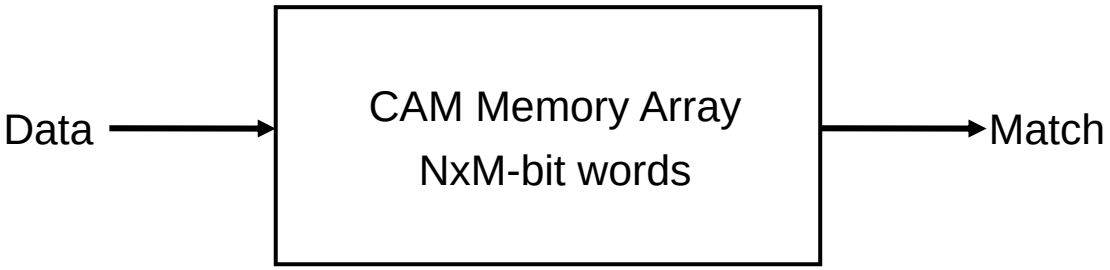
- Content addressable memories (CAMs) play an important role in digital systems and applications
  - such as communication, networking, data compression, etc.
- Each storage element of a CAM has the hardware capability to store and compare its contents with the data broadcasted by the control unit
- CAM types
  - dynamic or static
  - binary or ternary
- The binary static CAM is discussed

Difference Between RAM and CAM

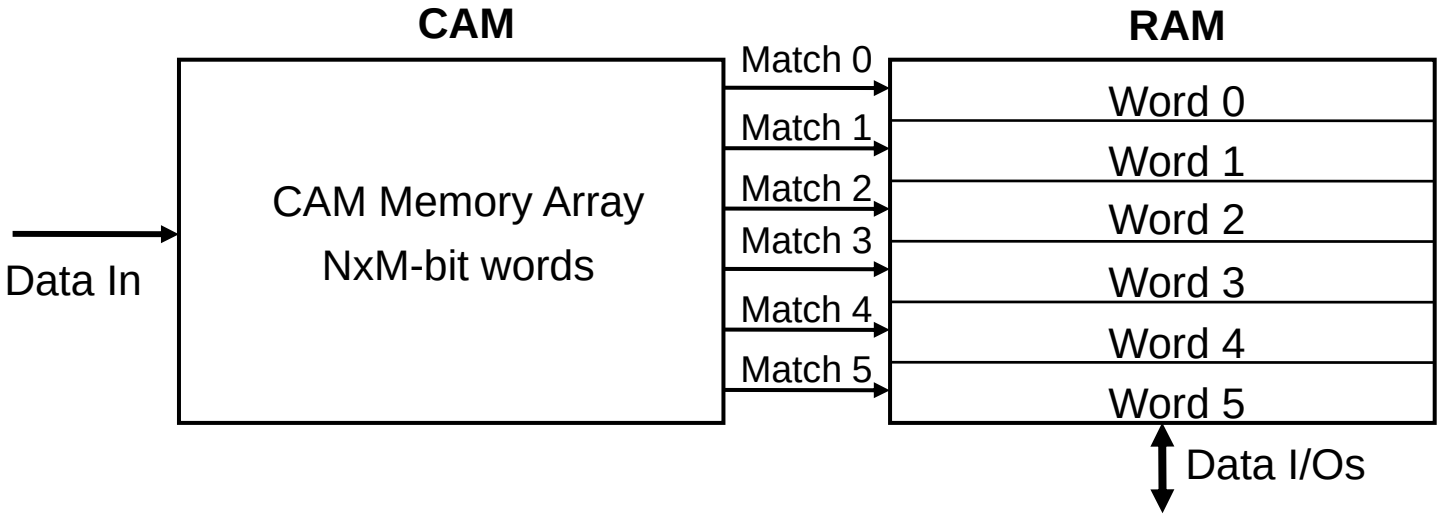


CAM – *Applications*

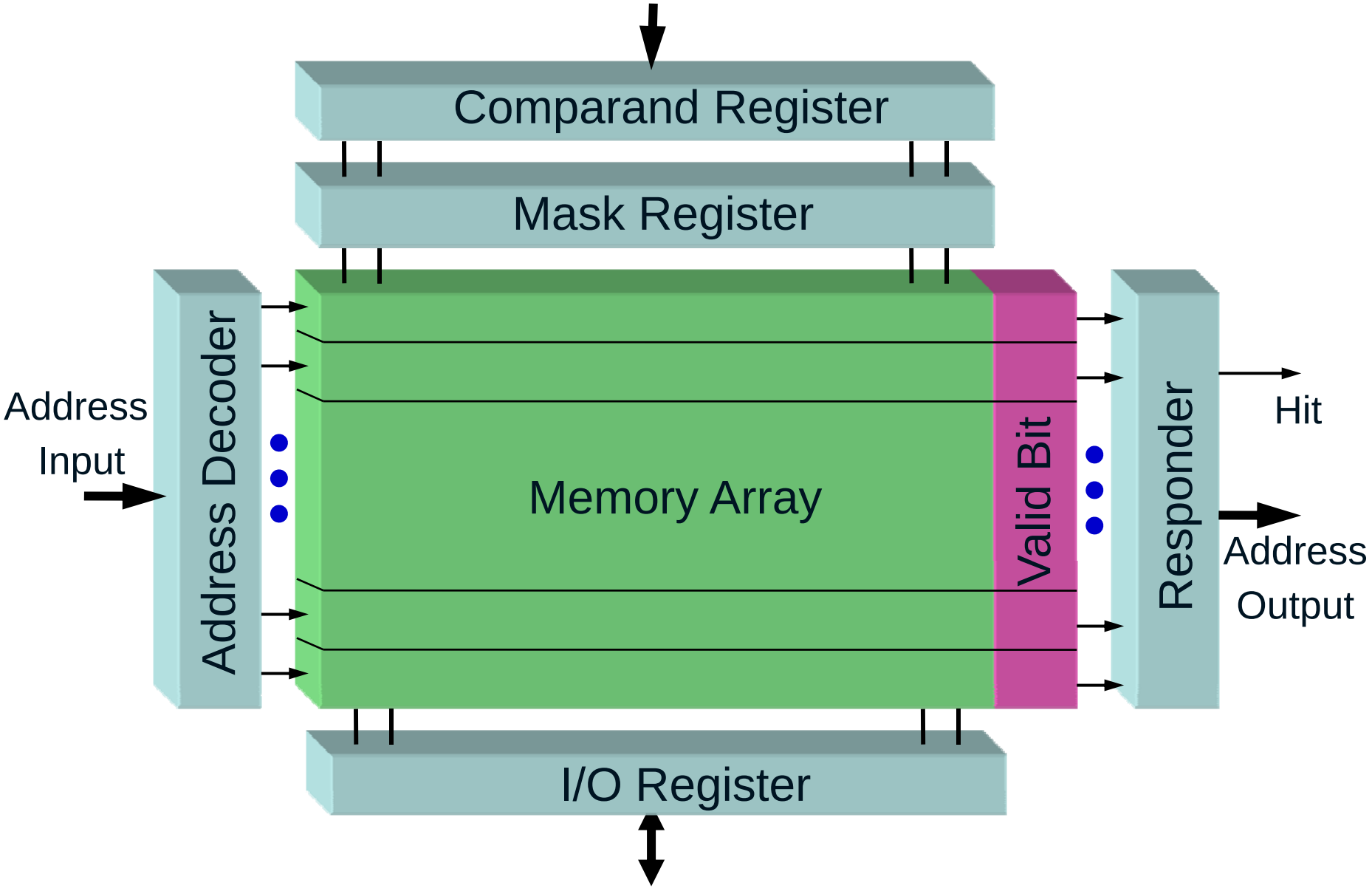
CAM architecture



Cache architecture



CAM – *Architecture*

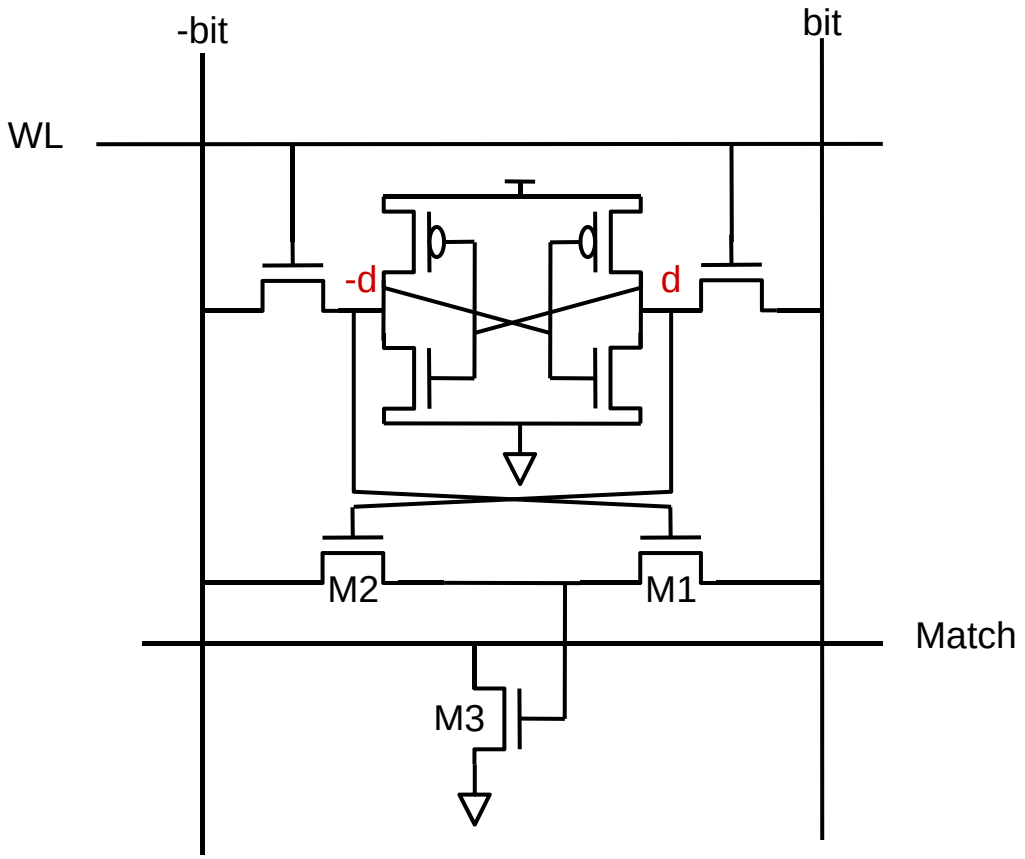


CAM – *Basic Components*

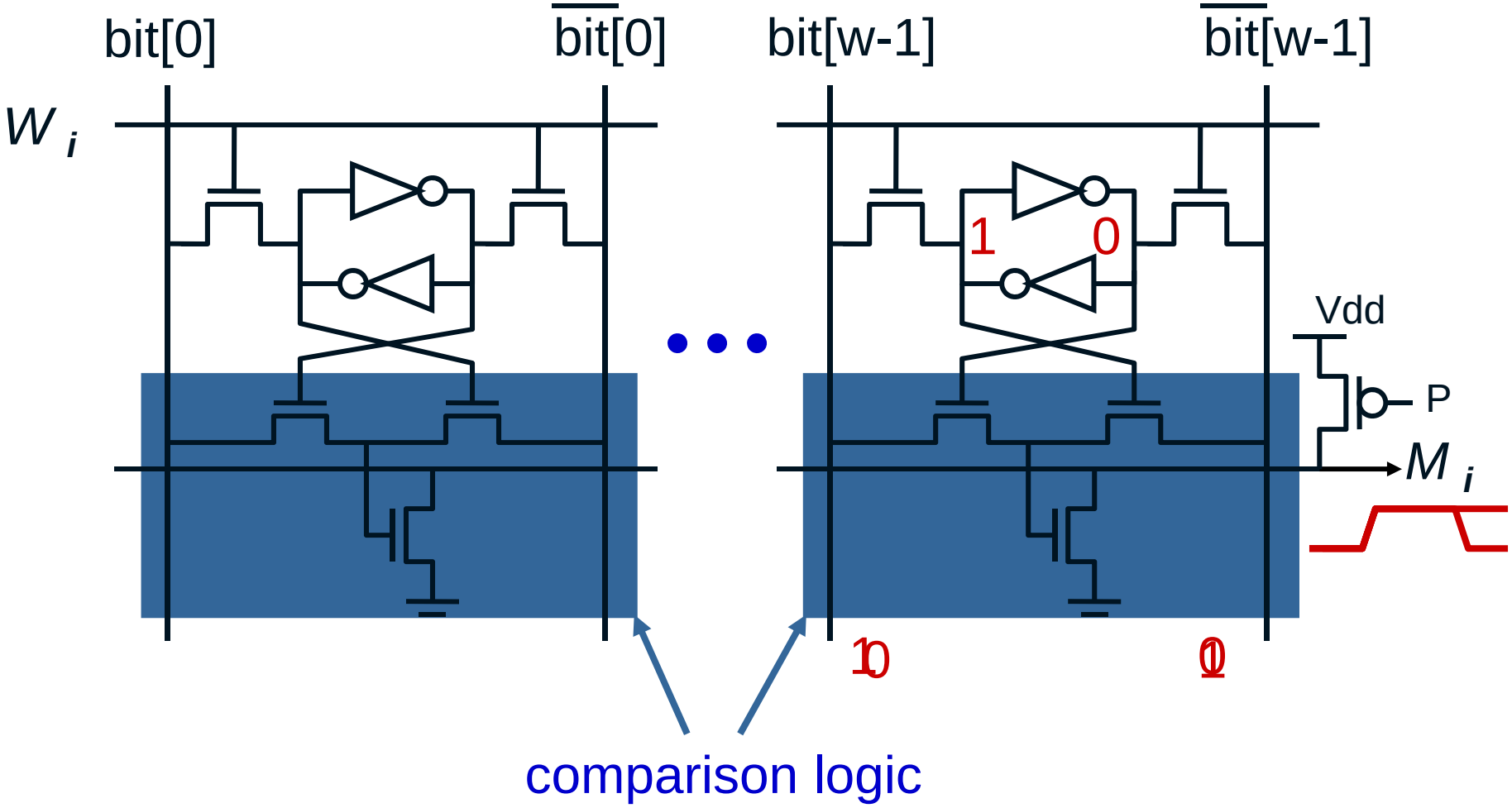
- Comparand Register
  - It contains the data to be compared with the content of the memory array
- Mask Register
  - It is used to mask off portions of the data word(s) which do not participate in the operations
- Memory Array
  - It provides storage and search (compare) medium for data
- Responder
  - It indicates success or failure of a compare operation

CAM – *Binary Cell*

CAM cell

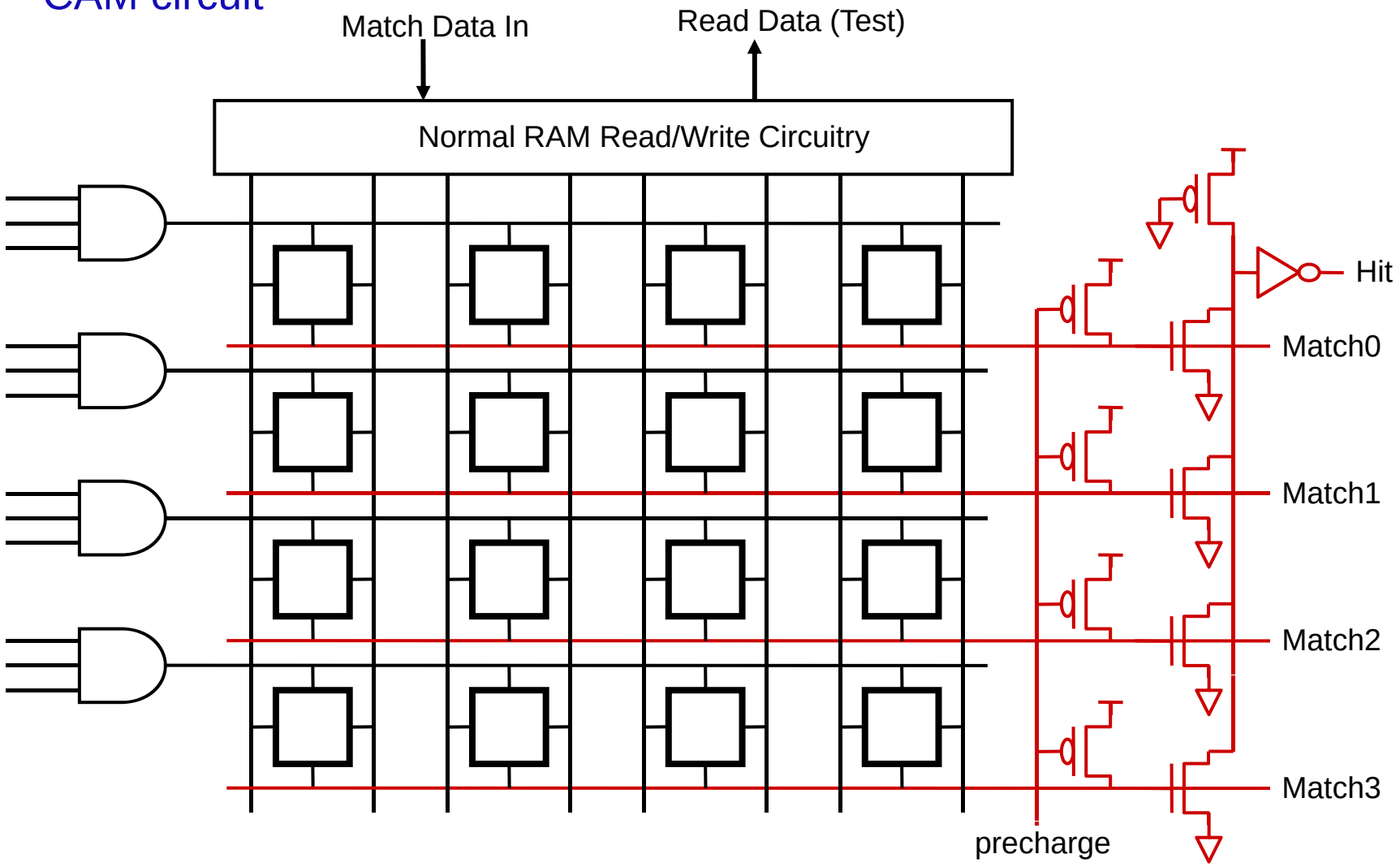


CAM – *Word Structure*



CAM – *Organization*

CAM circuit



Non-volatile Memory – *Flash Memory*

- Flash memory
  - *A nonvolatile, in-system-updateable, high-density memory technology that is per-bit programmable and per-block or per-chip erasable*
- In-system updateable
  - A memory whose contents can be easily modified by the system processor
- Block size
  - The number of cells that are erased at the same time
- Cycling
  - The process of programming and erasing a flash memory cell

Flash Memory – *Definition*

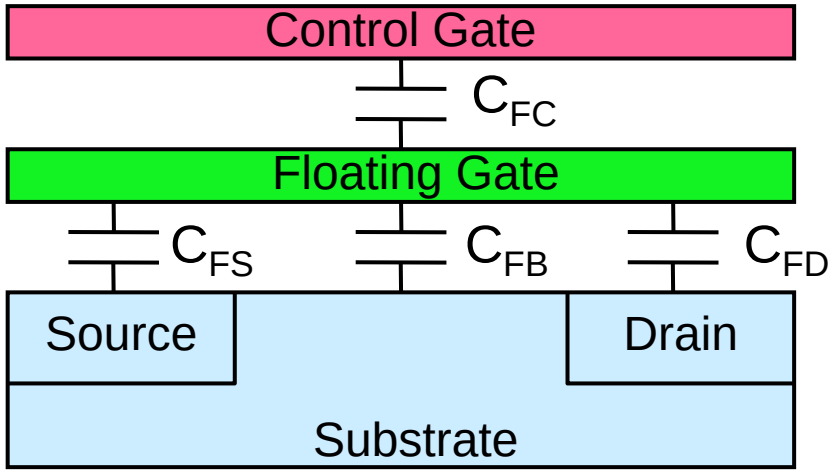
- Erase
  - To change a flash memory cell value *from 0 to 1*
- Program
  - To change a flash memory cell value *from 1 to 0*
- Endurance
  - The capability of maintaining the stored information after erase/program/read cycling
- Retention
  - The capability of keeping the stored information in time

Basics

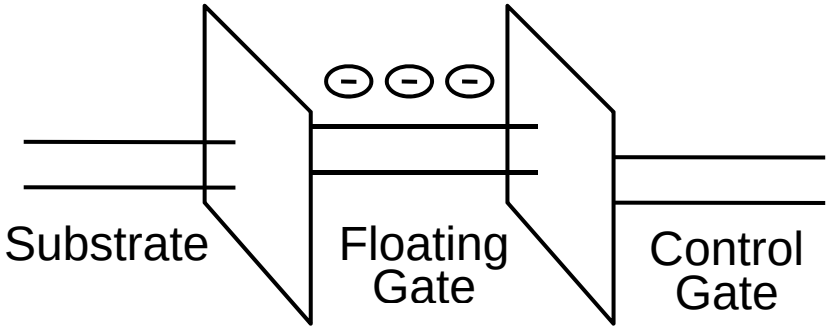
- How can a memory cell commute from one state to the others independently of external condition?
  - One solution is to have a transistor with a *threshold voltage that can change repetitively from a high to a low state*
  - The high state corresponding to the binary value "1"
  - The low state corresponding to the binary value "0"
- Threshold voltage of a MOS transistor
  - $V_T = K - Q'/C_{ox}$
  - K is a constant depending on the gate and substrate material, doping, and gate oxide thickness
  - Q' is the charge in the gate oxide
  - C_{ox} is the gate oxide capacitance

Floating Gate Transistor

- Floating gate (FG) transistor



- Energy band diagram of an FG transistor



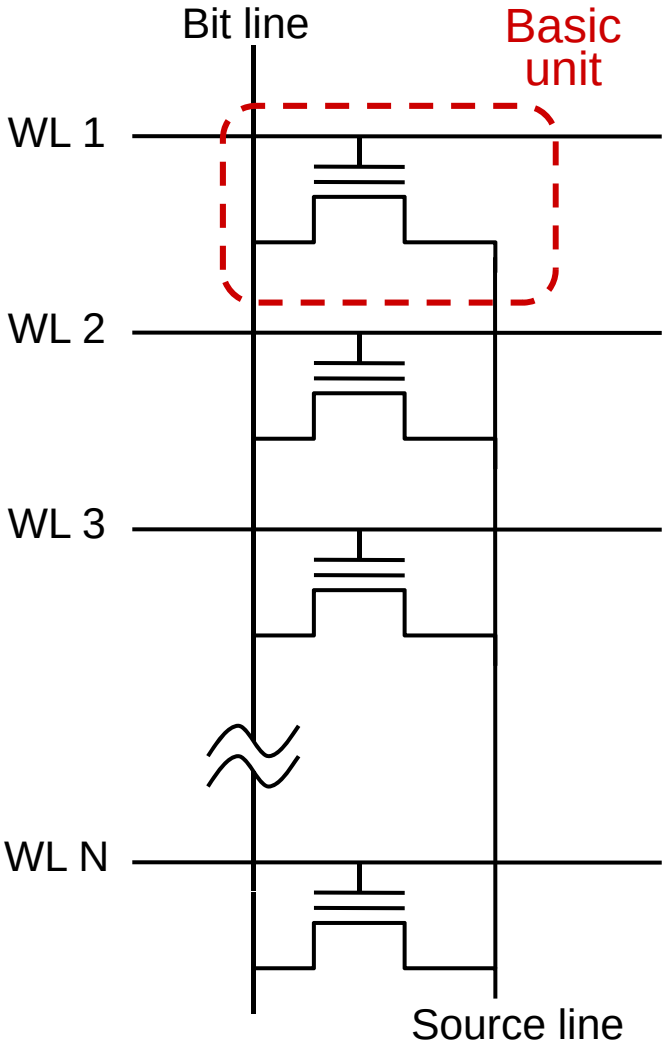
Flash Memory – *Threshold Voltage*

- When the voltages (V_{CG} & V_D) are applied to the control gate and the drain, the voltage at the floating gate (V_{FG}) by capacitive coupling is expressed as
  - $$V_{FG} = \frac{Q_{FG}}{C_{total}} + \frac{C_{FC}}{C_{total}} V_{CG} + \frac{C_{FD}}{C_{total}} V_D$$
  - $$C_{total} = C_{FC} + C_{FS} + C_{FB} + C_{FD}$$
- The minimum control gate voltage required to turn on the control gate is
  - $$V_T(CG) = \frac{C_{total}}{C_{FC}} V_T(FG) - \frac{Q_{FG}}{C_{FC}} - \frac{C_{FD}}{C_{FC}} V_D$$
  - where V_T(FG) is the threshold voltage to turn on the floating gate transistor
- The difference of threshold voltages between two memory data states (“0” and “1”) can be expressed as
  - $$\Delta V_T(CG) = -\frac{\Delta Q_{FG}}{C_{FC}}$$

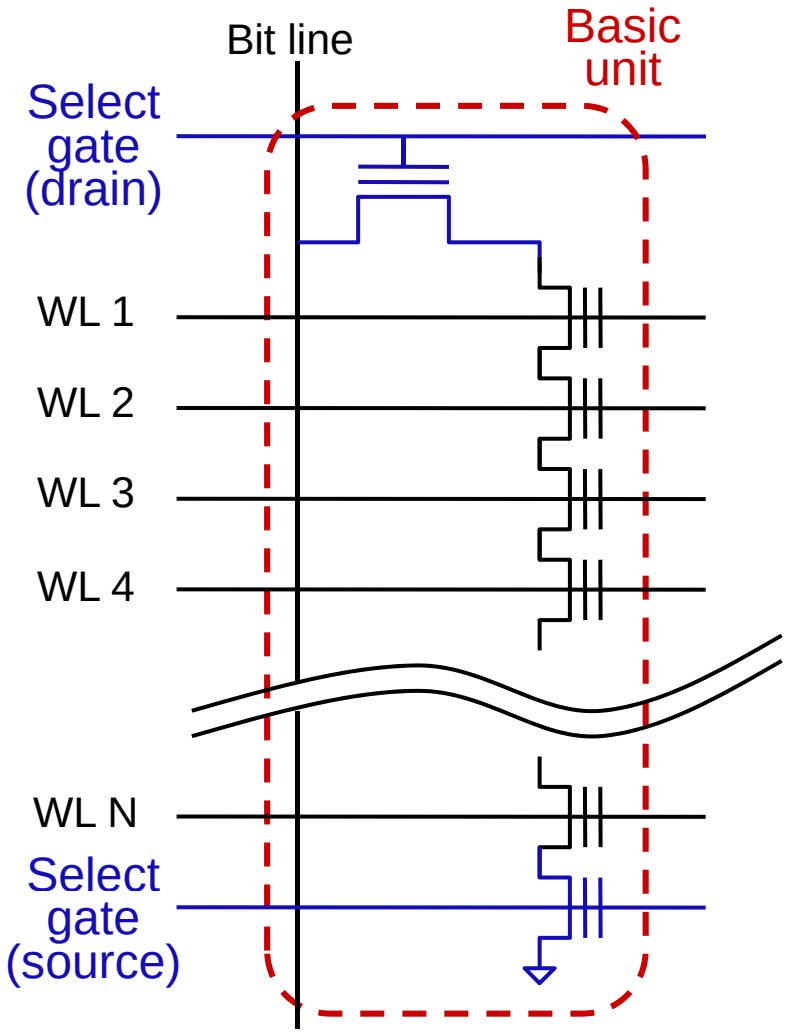
Flash Memory – *Structures*

- Two major flash memory structures
  - NOR & NAND
- NOR structure
  - Simplest
  - Dual power supply
  - Large block size
- NAND structure
  - Intermediate block size
  - High-speed and high density
  - For storage applications

Flash Memory – *Structures*



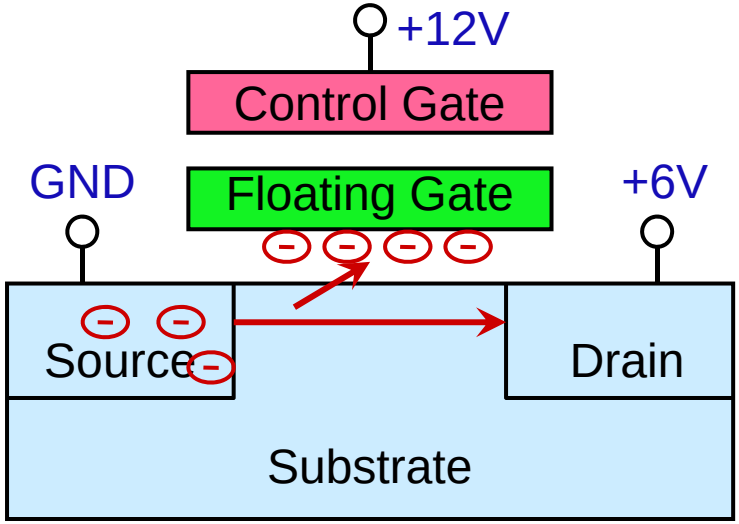
NOR structure



NAND structure

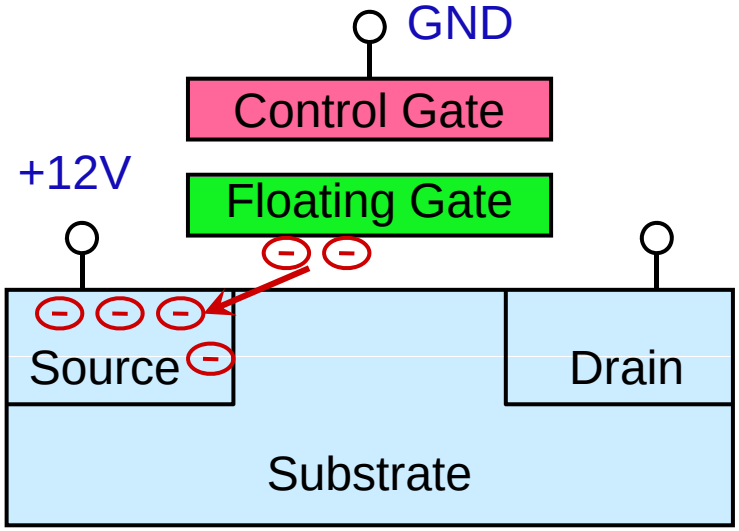
Flash Memory – *Program Operation*

- Program operation of the Intel's ETOX flash cells (NOR structure)
  - Apply 6V between drain and source
    - Generates hot electrons that are swept across the channel from source to drain
  - Apply 12 V between source and control gate
    - The high voltage on the control gate overcomes the oxide energy barrier, and attracts the electrons across the thin oxide, where they accumulate on the floating gate
  - Called channel hot-electron injection (HEI)



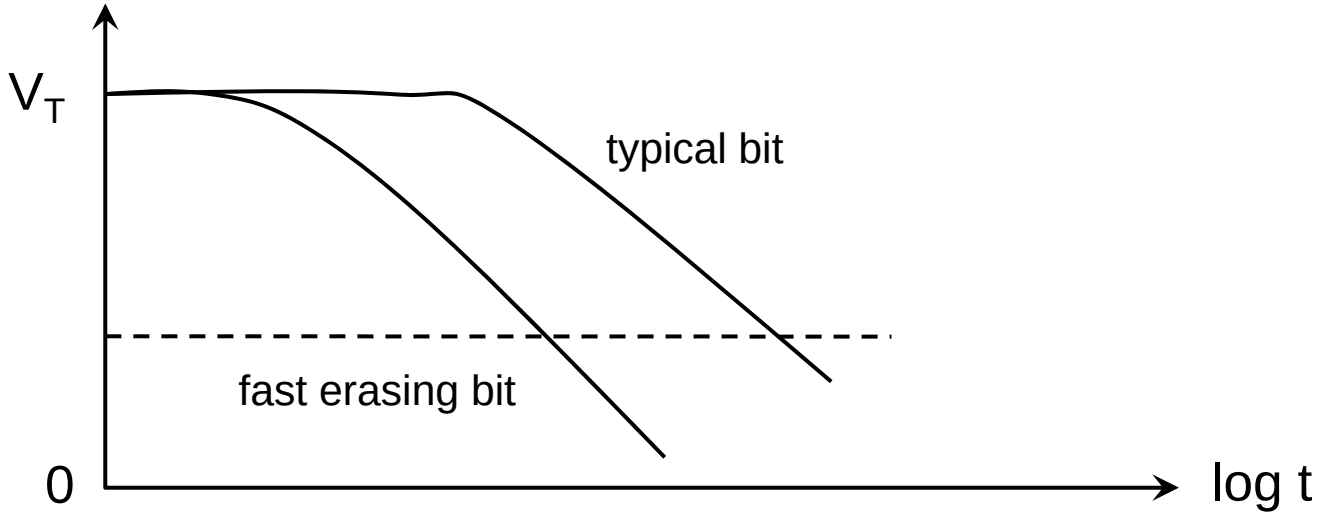
Flash Memory – *Erase Operation*

- Erase operation of the Intel's ETOX flash cells (NOR structure)
  - Floating the drain, grounding the control gate, and applying 12V to the source
  - A high electric field pulls electrons off the floating gate
  - Called Fowler-Nordheim (FN) tunneling



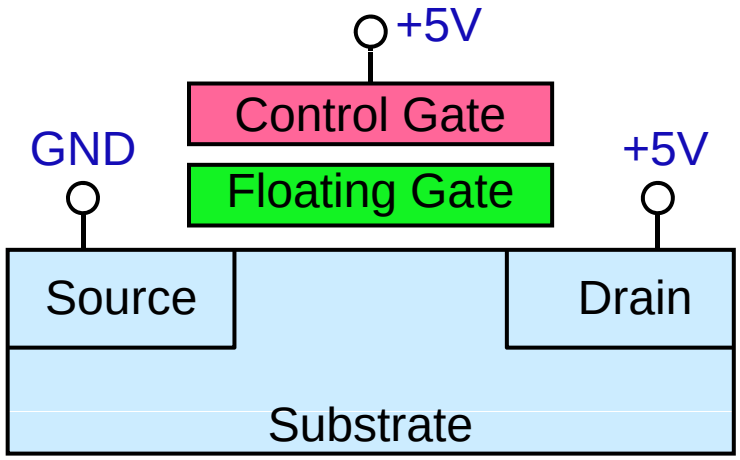
Flash Memory – *Erase Operation*

- Threshold voltage depends on oxide thickness
- Flash memory cells in the array may have slightly different gate oxide thickness, and the erase mechanism is not self-limiting
- After an erase pulse we may have "typical bits" and "fast erasing bits"

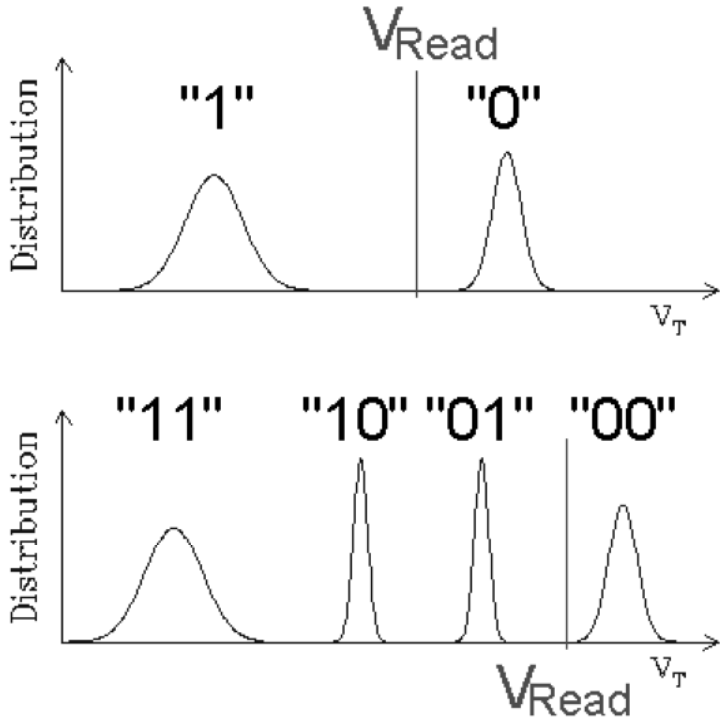


Flash Memory – *Read Operation*

- Read operation of the Intel's ETOX flash cells (NOR structure)
  - Apply 5V on the control gate and drain, and source is grounded
  - In an erased cell, the $V_C > V_T$
    - The drain to source current is detected by the sense amplifier
  - In a programmed cell, the $V_C < V_T$
    - The applied voltage on the control gate is not sufficient to turn it on. The absence of current results in a 0 at the corresponding flash memory output



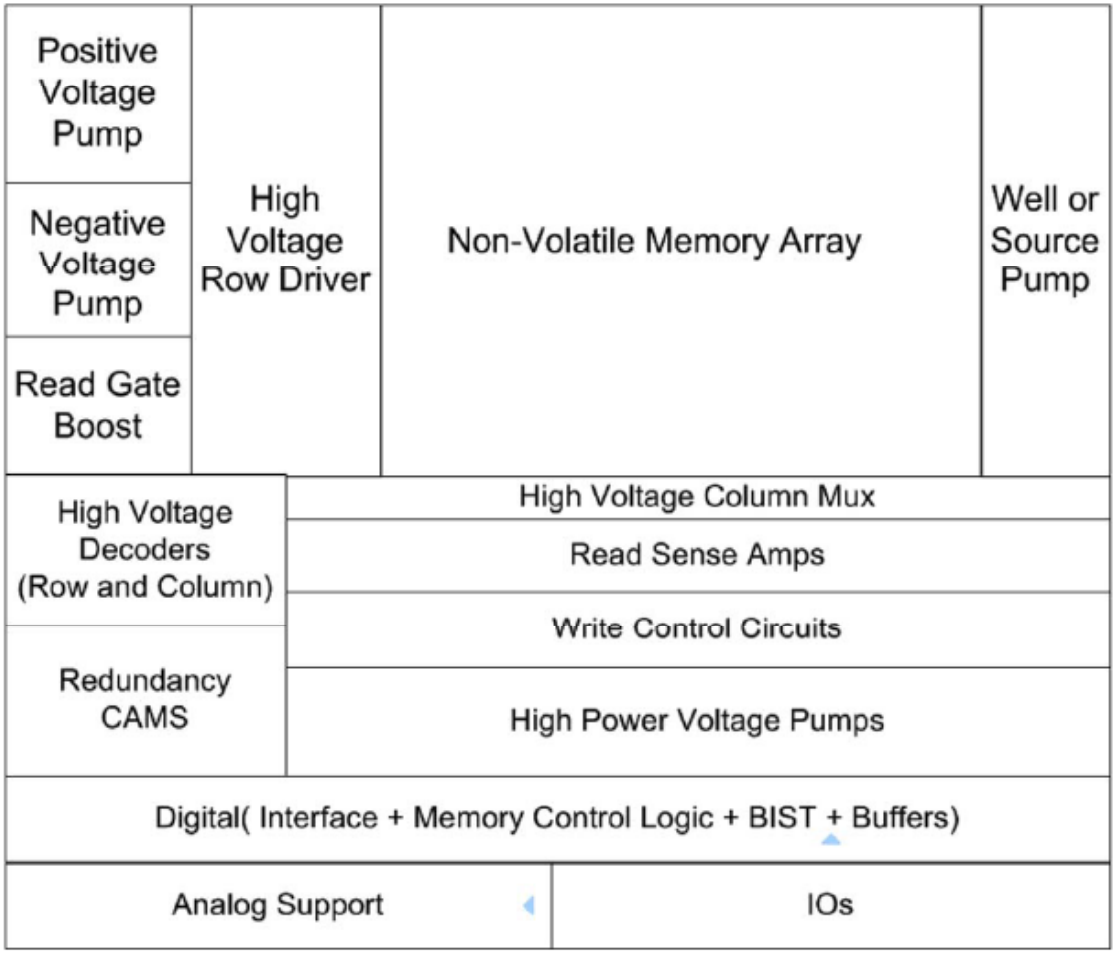
Flash Memory –*Concept of Multi-Level Flash Memories*



- 1 bit/cell => 2 levels
- 2 bit/cell => 4 levels

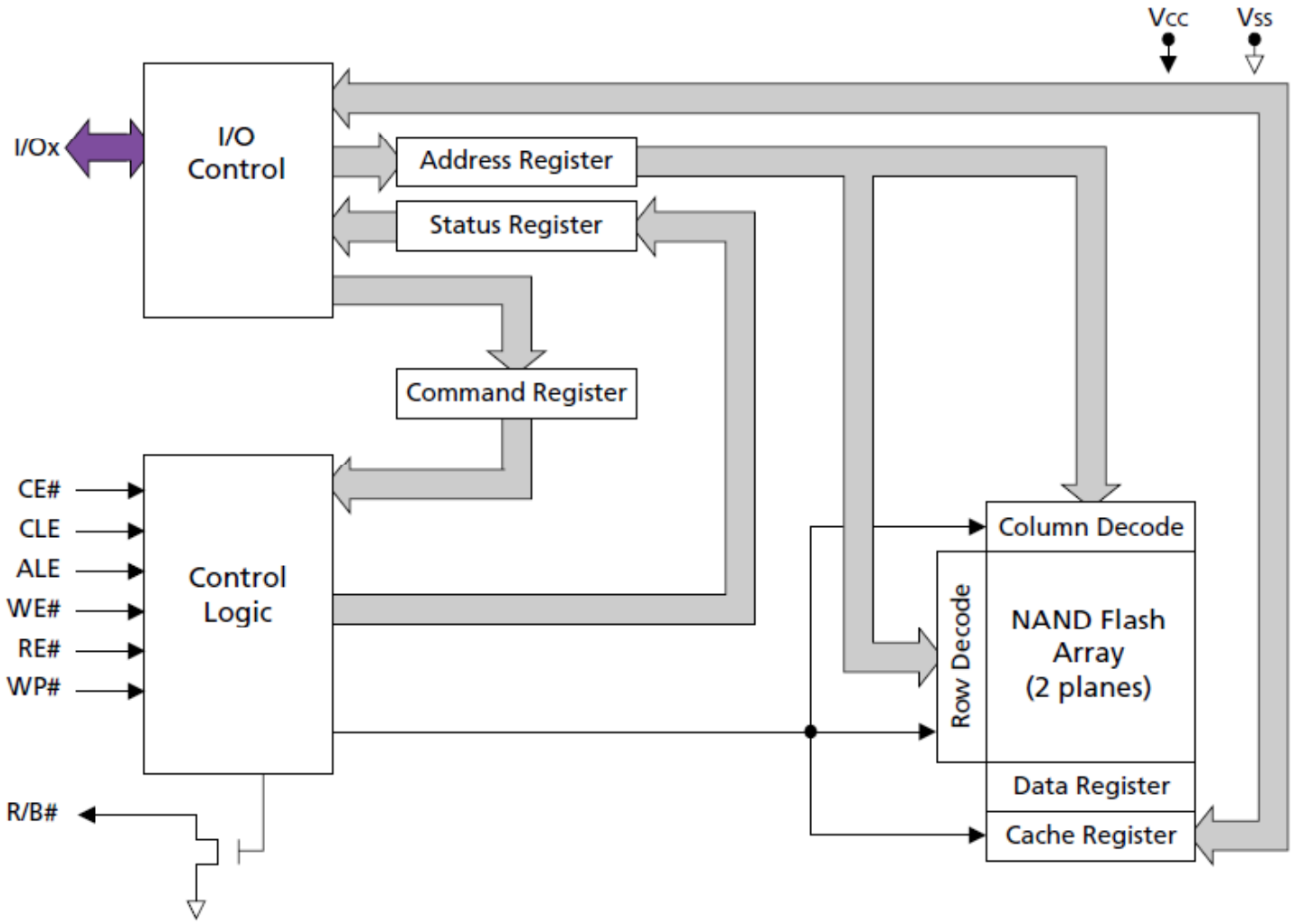
Source: Proceedings of IEEE, 2003

Flash Memory – *Basic Building Blocks of NOR Flash*



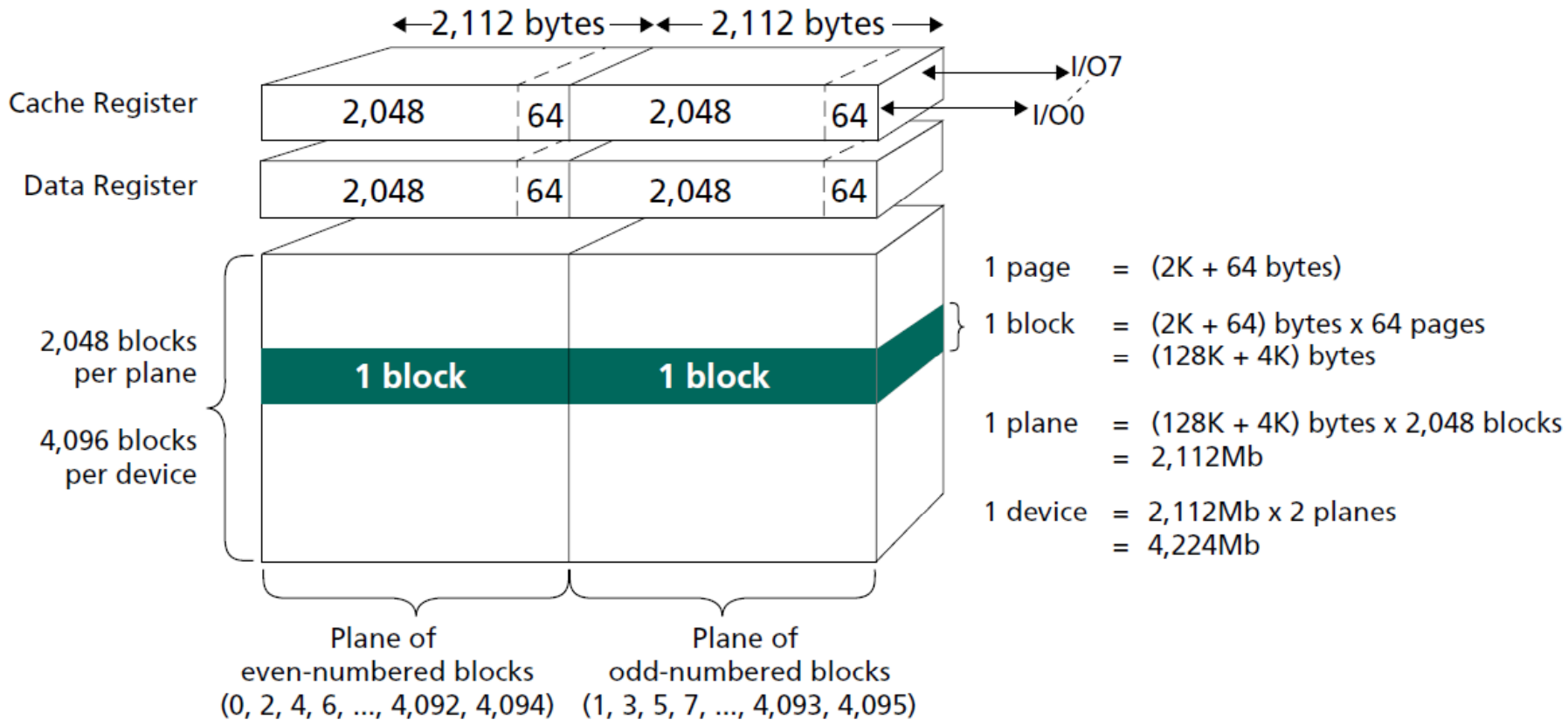
Source: Proceedings of IEEE, 2010

Flash Memory – *NAND Flash Functional Block Diagram*



Source: Micron

Flash Memory – *NAND Flash Array Organization*



Source: Micron