Remaining Topics

- Semiconductor Memory
  - Reference: ‘Analysis and Design of Digital Integrated Circuits’ by Hodges, Jackson, and Saleh
- Digital Logic using BJTs
- Layout

Semiconductor Memories

Classification

	Volatile	Speed	Density
SRAM	Yes	High	Moderate
DRAM	Yes	Moderate	High
ROM	No	High	Moderate
EPROM	No	Low	Low
Flash	No	Low	High
MRAM	No	High	Moderate
FeRAM	No	High	Moderate

Classification

	Volatile	Speed	Density
→ SRAM	Yes	High	Moderate
→ DRAM	Yes	Moderate	High
ROM	No	High	Moderate
EPROM	No	Low	Low
→ Flash	No	Low	High
MRAM	No	High	Moderate
FeRAM	No	High	Moderate

Memory Classification

Stand Alone Memory
- DRAM, ROM, SRAM, FLASH, EPROM
- MRAM, FeRAM

Embedded Memory
- SRAM for cache
- DRAM
- ROM

Memory Market

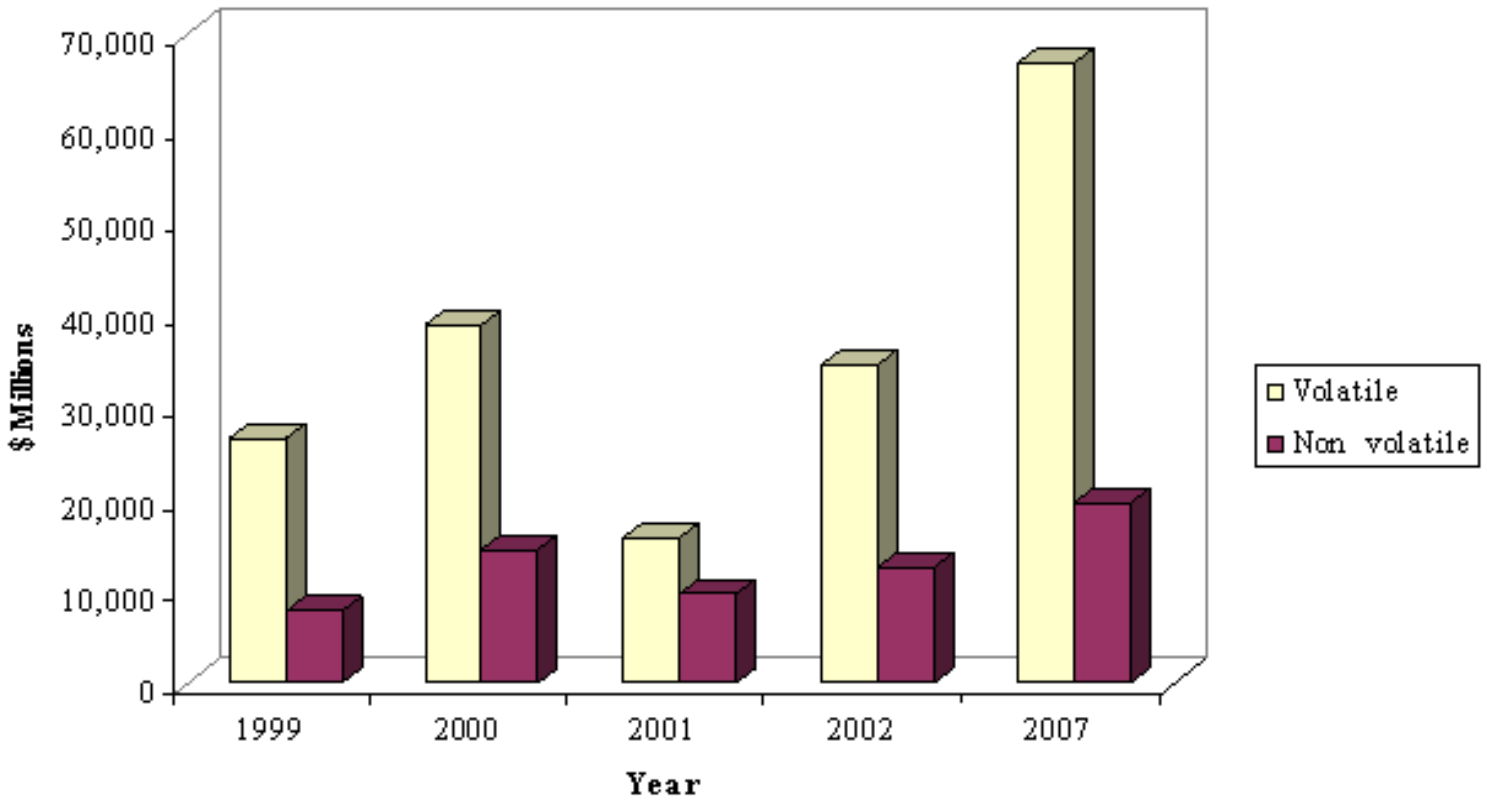
September 2006 Chip Sales Reached $21.4 Billion

SAN JOSE, CA – November 16, 2006 – The Semiconductor Industry Association (SIA) today released its annual forecast of global semiconductor sales, projecting that the industry will continue to ride a strong wave of consumer demand for electronic products, driving sales to $321 billion in 2009. The SIA forecast projects a compound annual growth rate of 9 percent for the forecast period, 2006-2009. Total worldwide microchip sales in 2005 amounted to $227.5 billion.

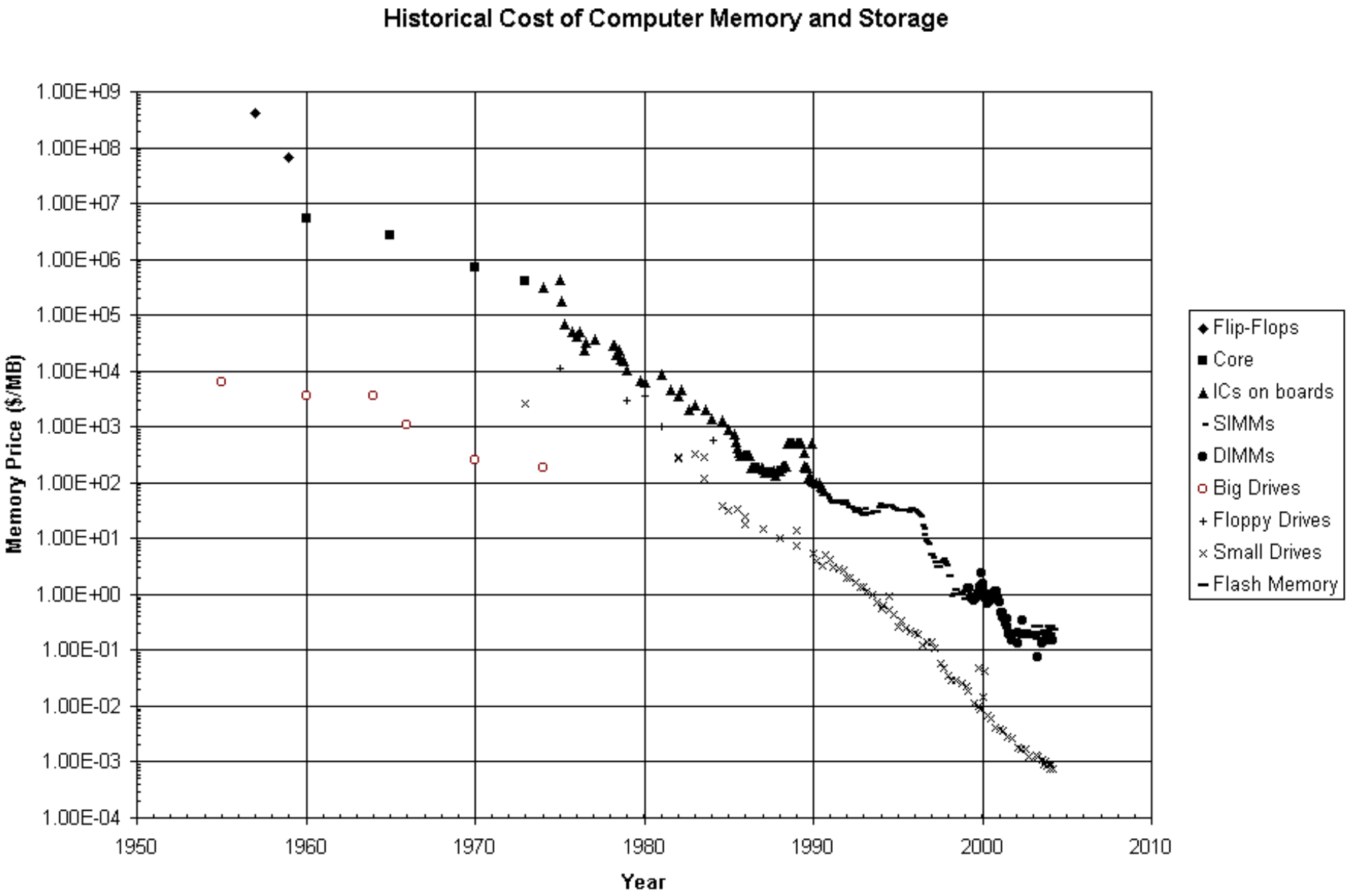
MEMORY accounted for ~25% of the total market
DRAM sales are projected to be the fastest-growing segment
- Compound annual growth rate >14% in 2006-2009
- Sales especially strong for remainder of 2006 and in 2007 --> Vista OS

Sales of NAND flash are projected to grow at compound annual rate of 11% through 2009.
- Driven by replacement of hard drives; Laptop computers with pure NAND drives are projected to account for up to 25 percent of the laptop market by 2009

Memory Market

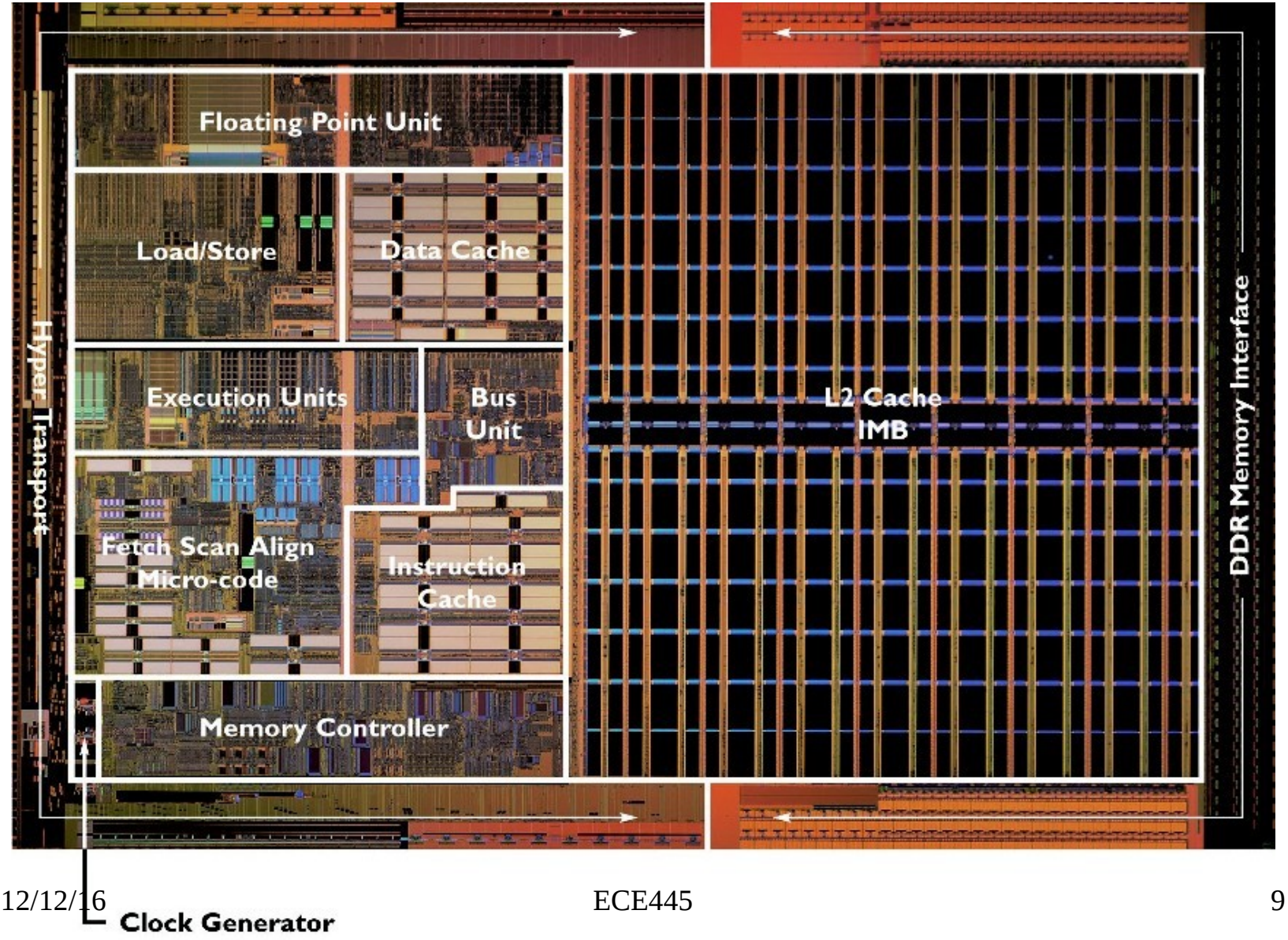


Source: BCC, Incorporated

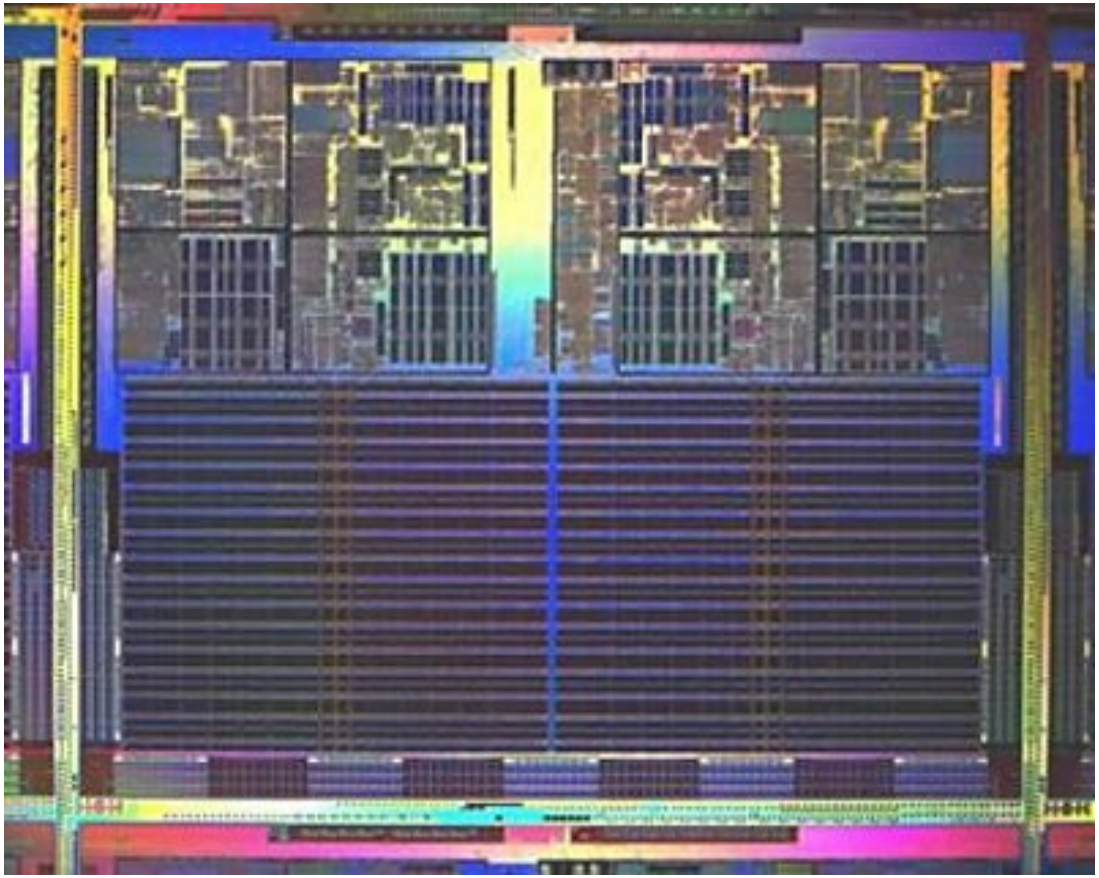


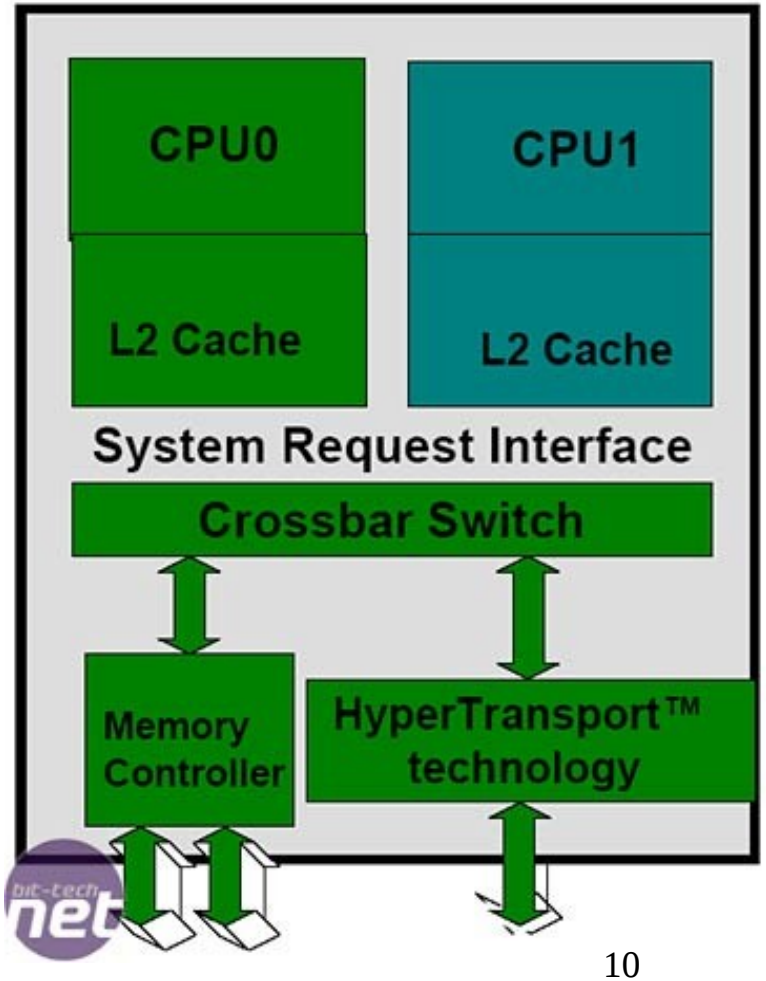
http://www.jcmit.com/mem2002.htm

AMD Athlon


Floating Point Unit
Load/Store
Data Cache
Execution Units
Bus Unit
Fetch Scan Align Micro-code
Instruction Cache
Memory Controller
L2 Cache 1MB
HyperTransport
DDR Memory Interface
Clock Generator

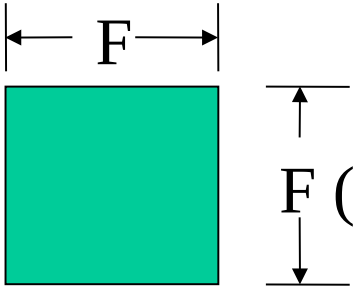
AMD Dual-Core Athlon

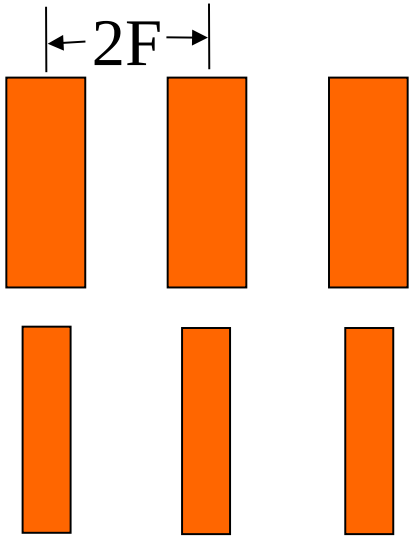




Memory Density

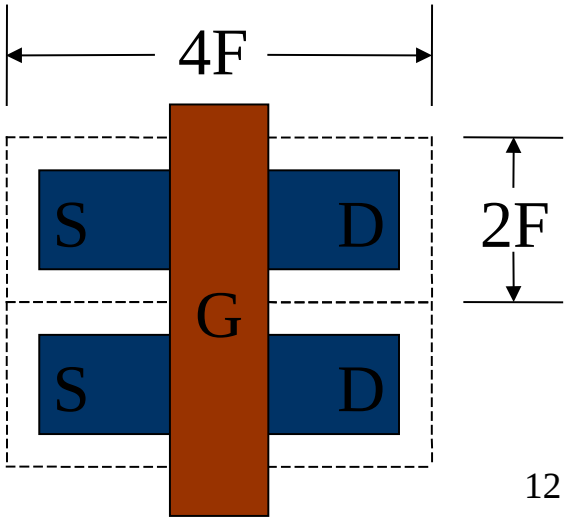
Cell Size



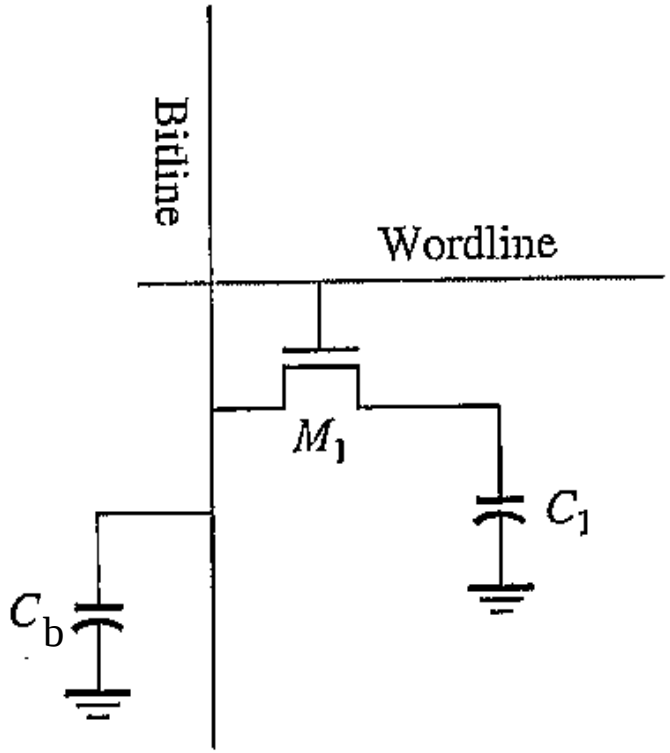


SRAM 6-T cell:
40-100 F^2 cell

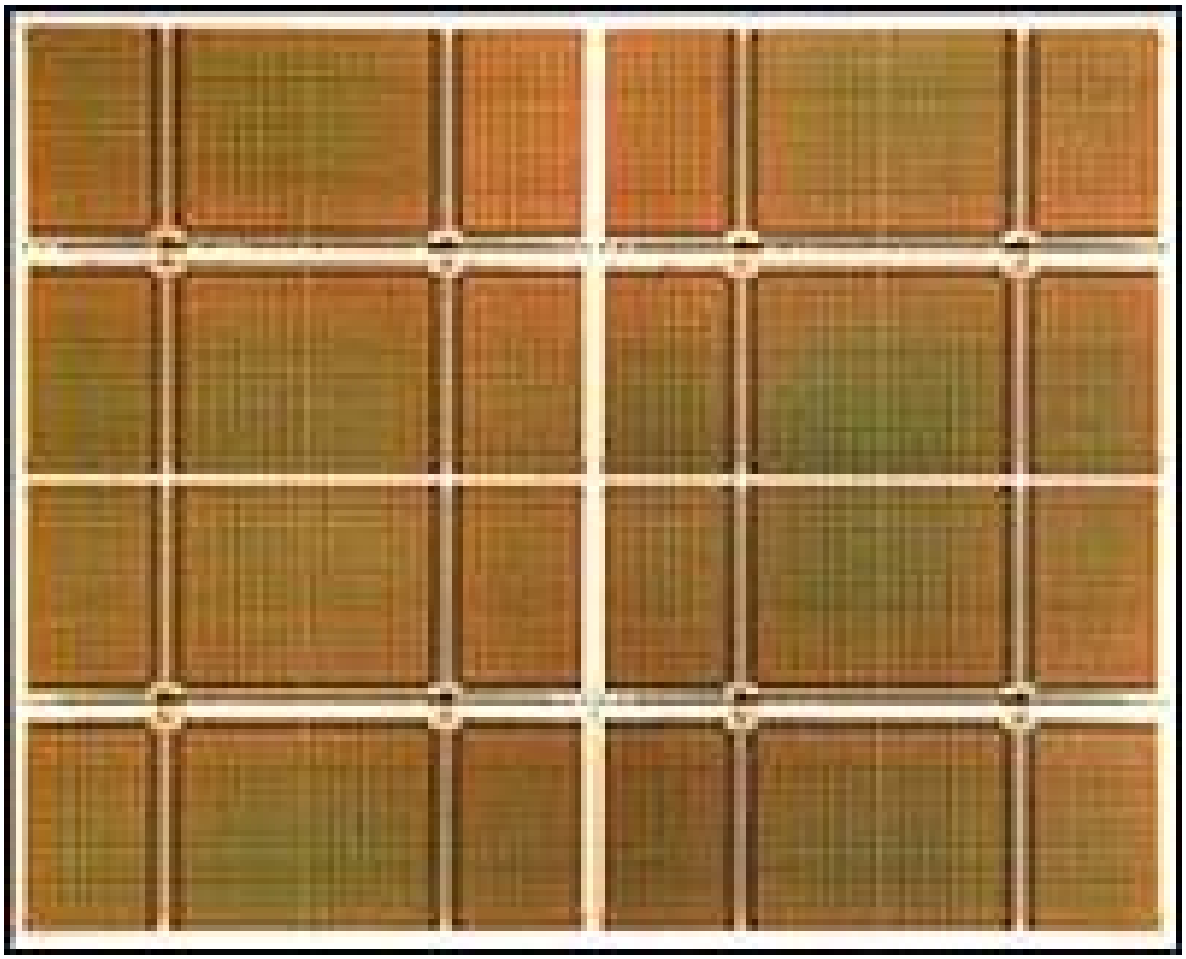
DRAM 1-T, 1-C cell:
5-10 F^2 cell; most common $8F^2$ cell



1T DRAM Cell

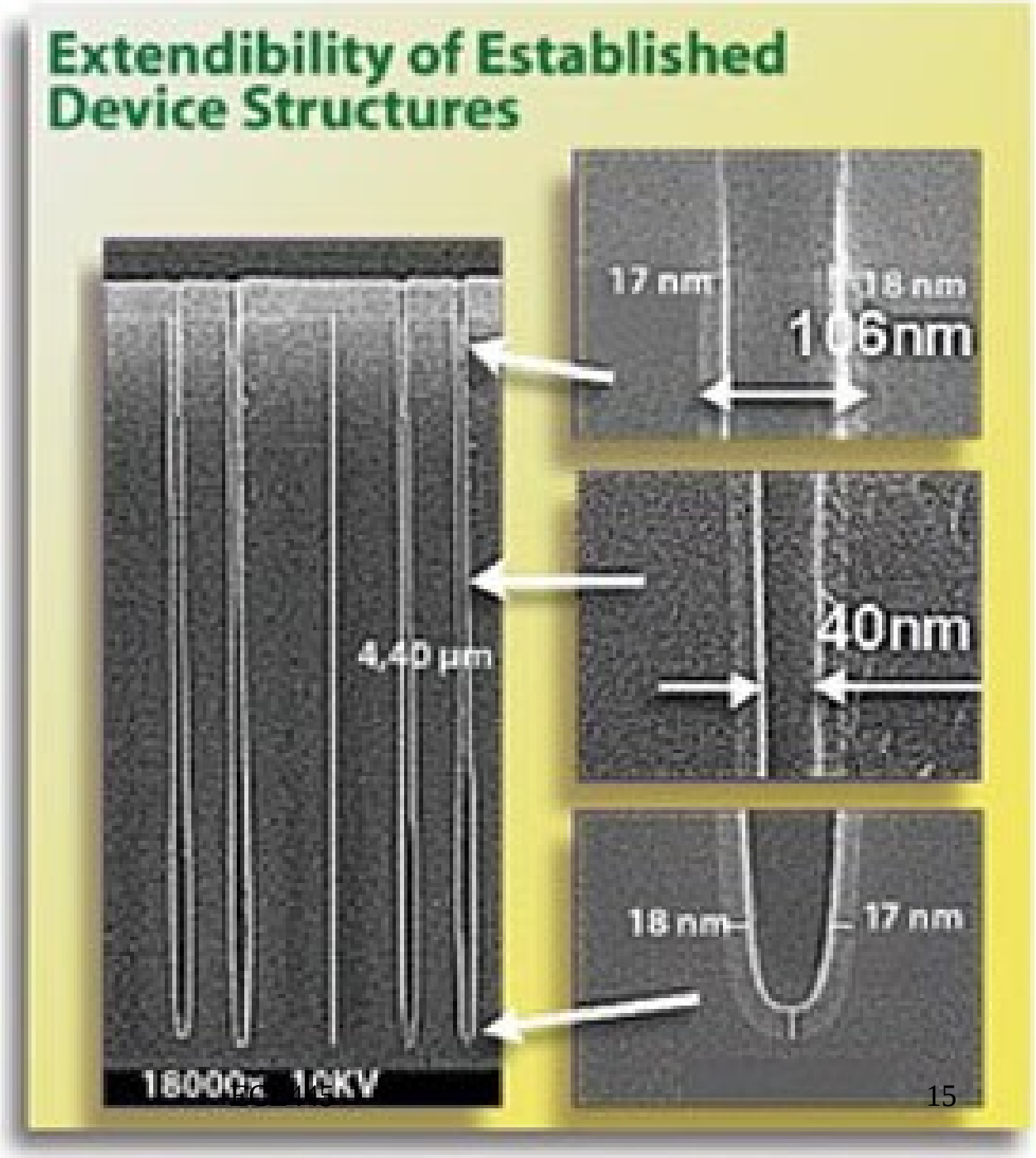


4Gb DRAM



Infineon 2004

DRAM Trench Capacitor



Redundancy

Redundancy is always used on standalone DRAMs and frequently for embedded DRAMs.

Implemented via on chip fuses in which fuses are blown either by a laser pulse or by an electrical pulse during wafer test.

Normally a few rows and columns of redundancy are offered on a large DRAM embedded macro. Smaller macros require less redundancy (or no redundancy)

Redundancy can also be implemented to some extent in the logic part of the circuits. To implement a spare decoder, for example, switches can be used that shift all decoder connections by one to bypass the faulty decoder.

Memory Standards

Most memory standards are controlled by the **JEDEC** Solid State Technology Association (Once known as the **J**oint **E**lectron **D**evice **E**ngineering **C**ouncil). JEDEC is the semiconductor engineering standardization body of the Electronic Industries Alliance (EIA), a trade association that represents all areas of the electronics industry and was originally created in 1960 to cover the standardization of discrete semiconductor devices and later expanded in 1970 to include integrated circuits.

JEDEC does its work through its 48 committees/ subcommittees that are overseen by the **JEDEC** Board of Directors. Presently there are about 300 member companies in **JEDEC** including both manufacturers and users of semiconductor components and others allied to the field.

DRAM Architecture

More than 10 Different Types of DRAM

Page Mode DRAM	DRAM
Fast Page Mode DRAM	FPM DRAM
Extended Data Output DRAM	EDO DRAM
Burse Extended Data Output DRAM	BEDO DRAM
Enhanced DRAM	EDRAM
Synchronous DRAM	SDRAM
PC100 Synchronous DRAM	PC100 SDRAM
Enhanced Synchronous DRAM	ESDRAM
Double Data Rate DRAM	DDR DRAM
Direct Rambus DRAM	DRDRAM
Synchronous Link DRAM	SLDRAM

…..

Memory Circuits

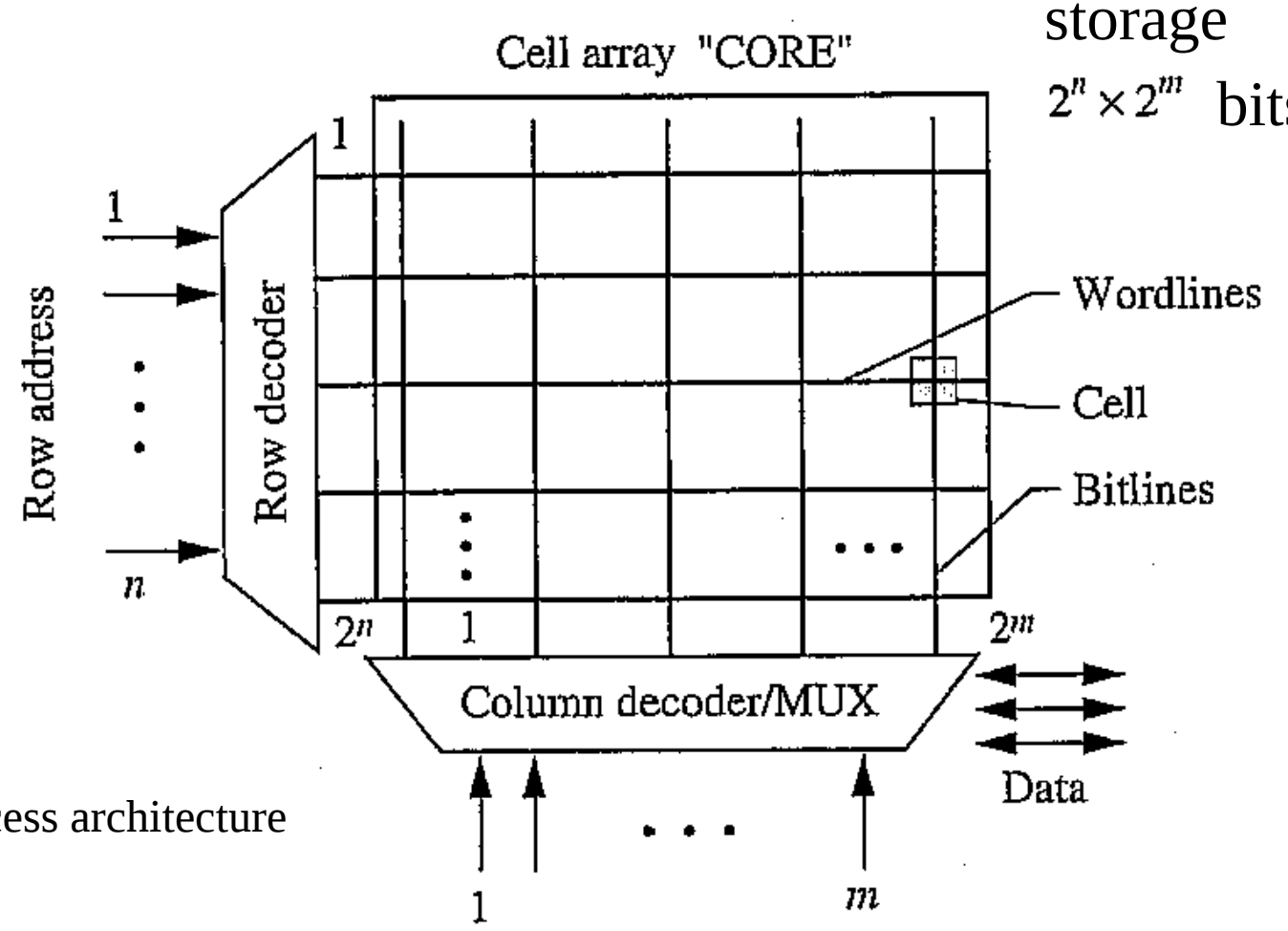
Memory Core - typically stores 1-bit
- can have multilevel cores)

Peripheral Circuits
- row and column decoders
- sense amplifiers
- column precharge circuits
- data buffers
- read/write (R/W) circuits
- …

Typical Memory Organization

storage

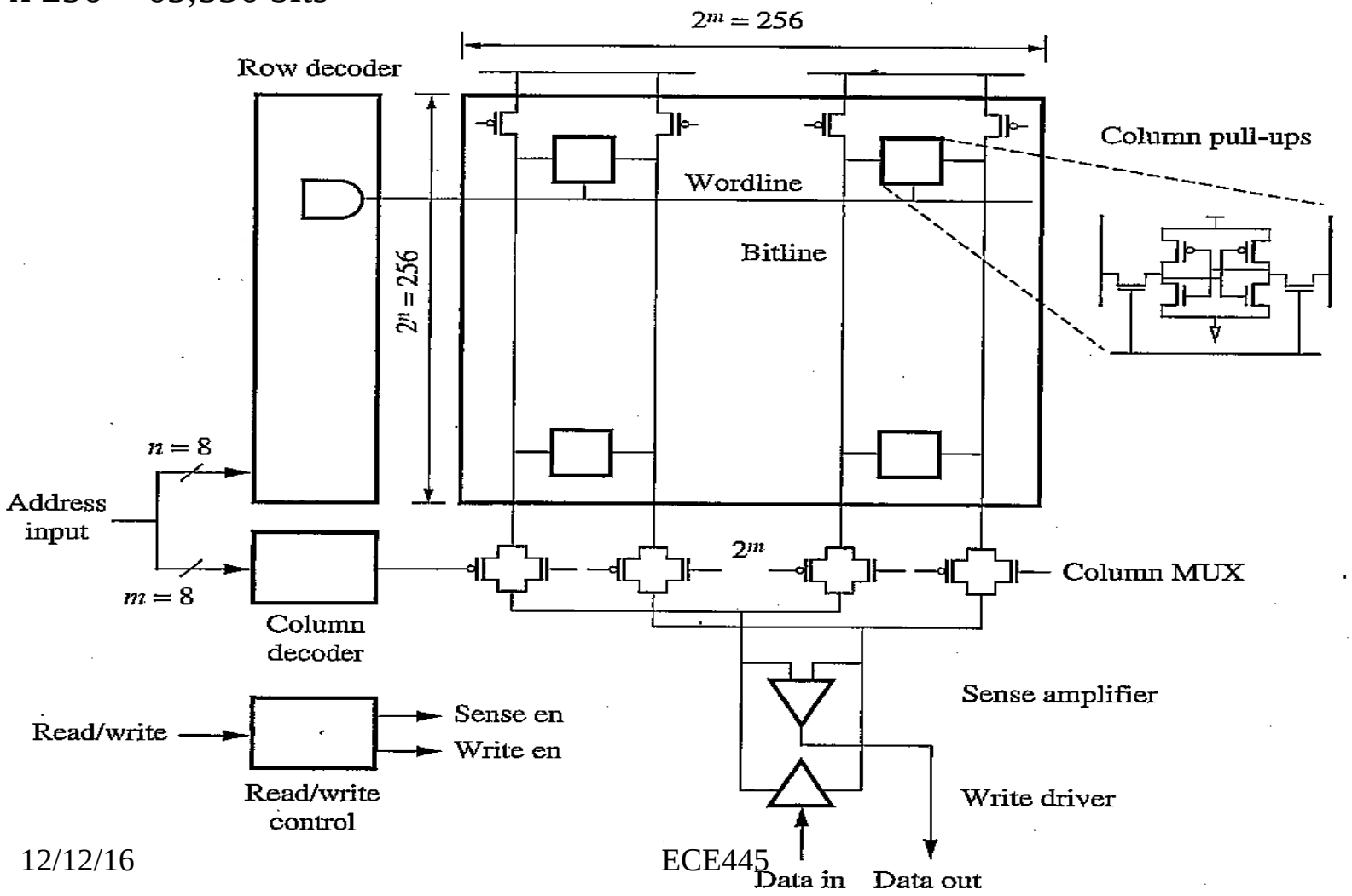
$2^n \times 2^m$ bits



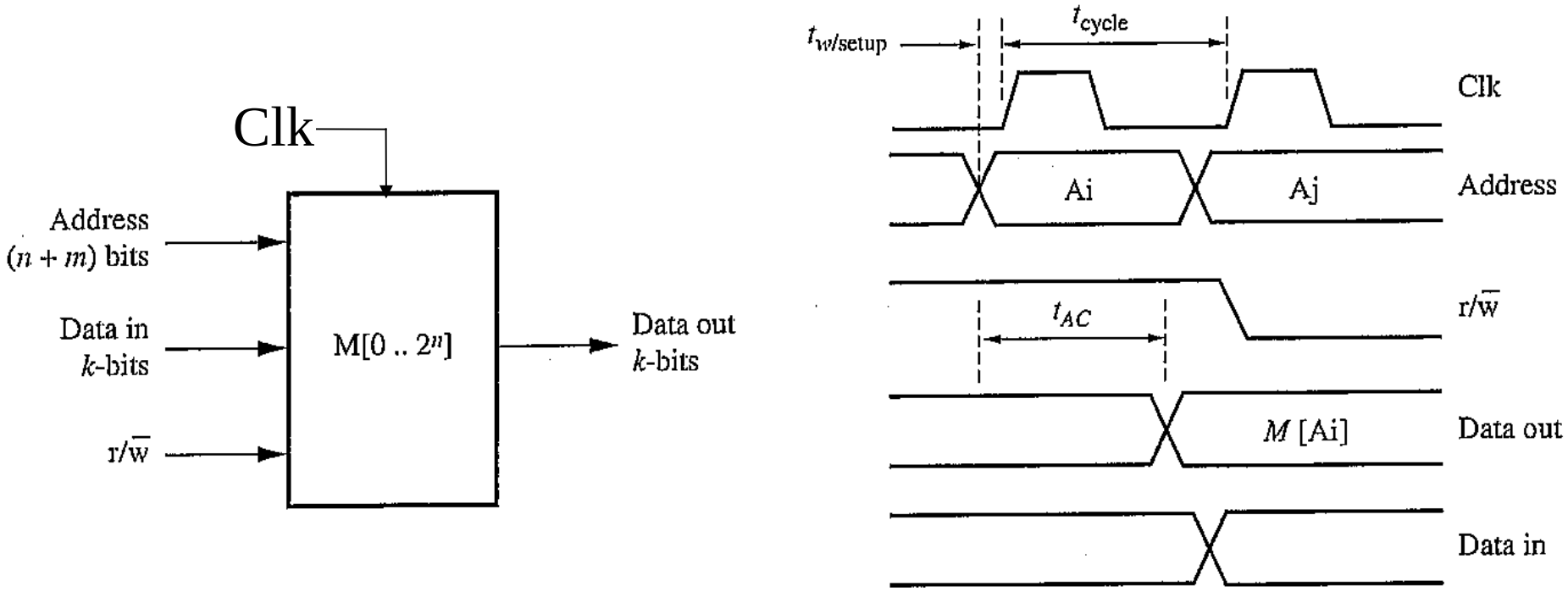
random access architecture

64 Kb RAM (2^8 x 2^8)

256 x 256 = 65,536 bits



Timing Signals

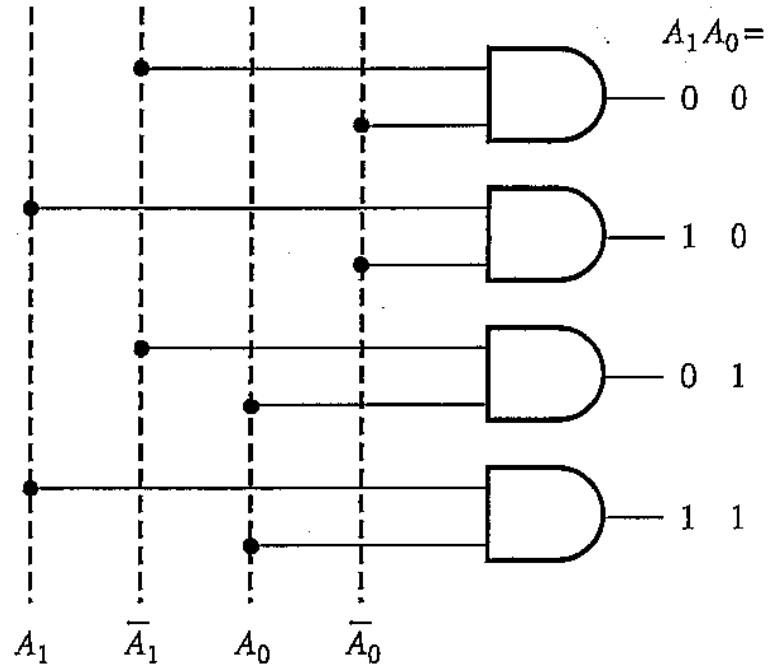


t_{CYCLE} – minimum time to complete successive read or write operations

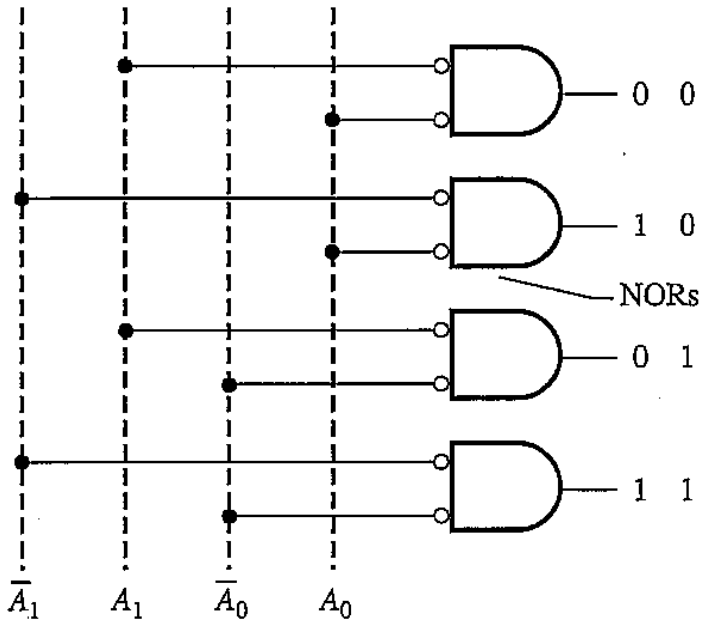
$t_{W/SETUP}$ – write setup time

t_{AC} – read access time – delay from change in the address until data from that address are available at the output

Row Decoder (n=2)



(a) AND-based decoder



(b) NOR-based decoder

Cascade Decoder

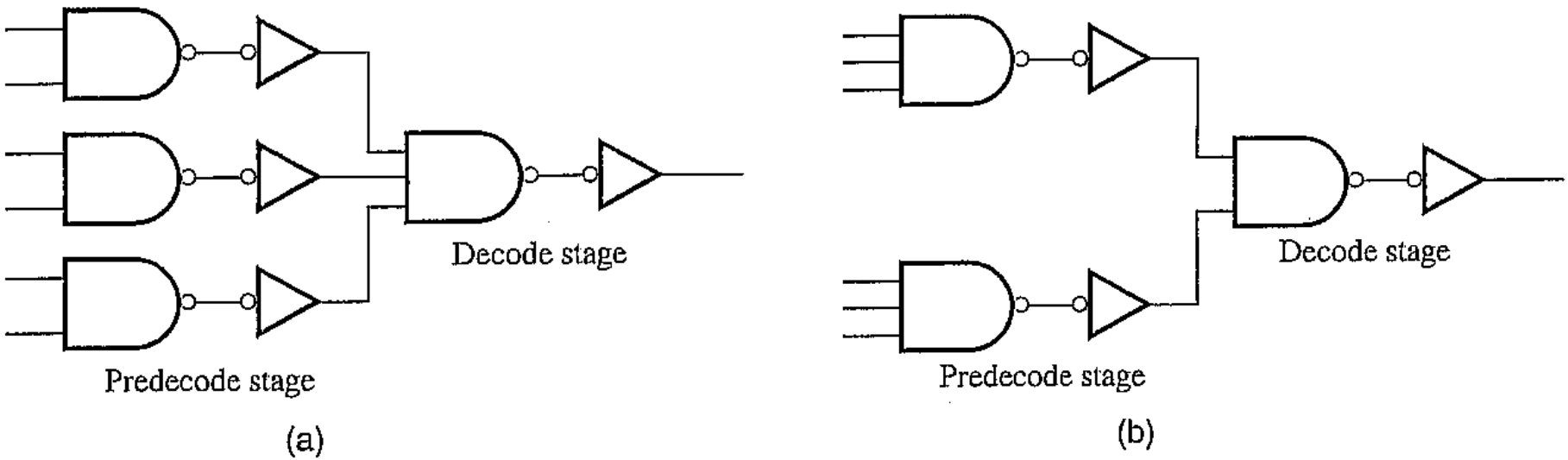
An n-bit decoder → 2^n logic gates each with n inputs

For n=6 → requires 64, 6–input NAND gates and 64 Inverters

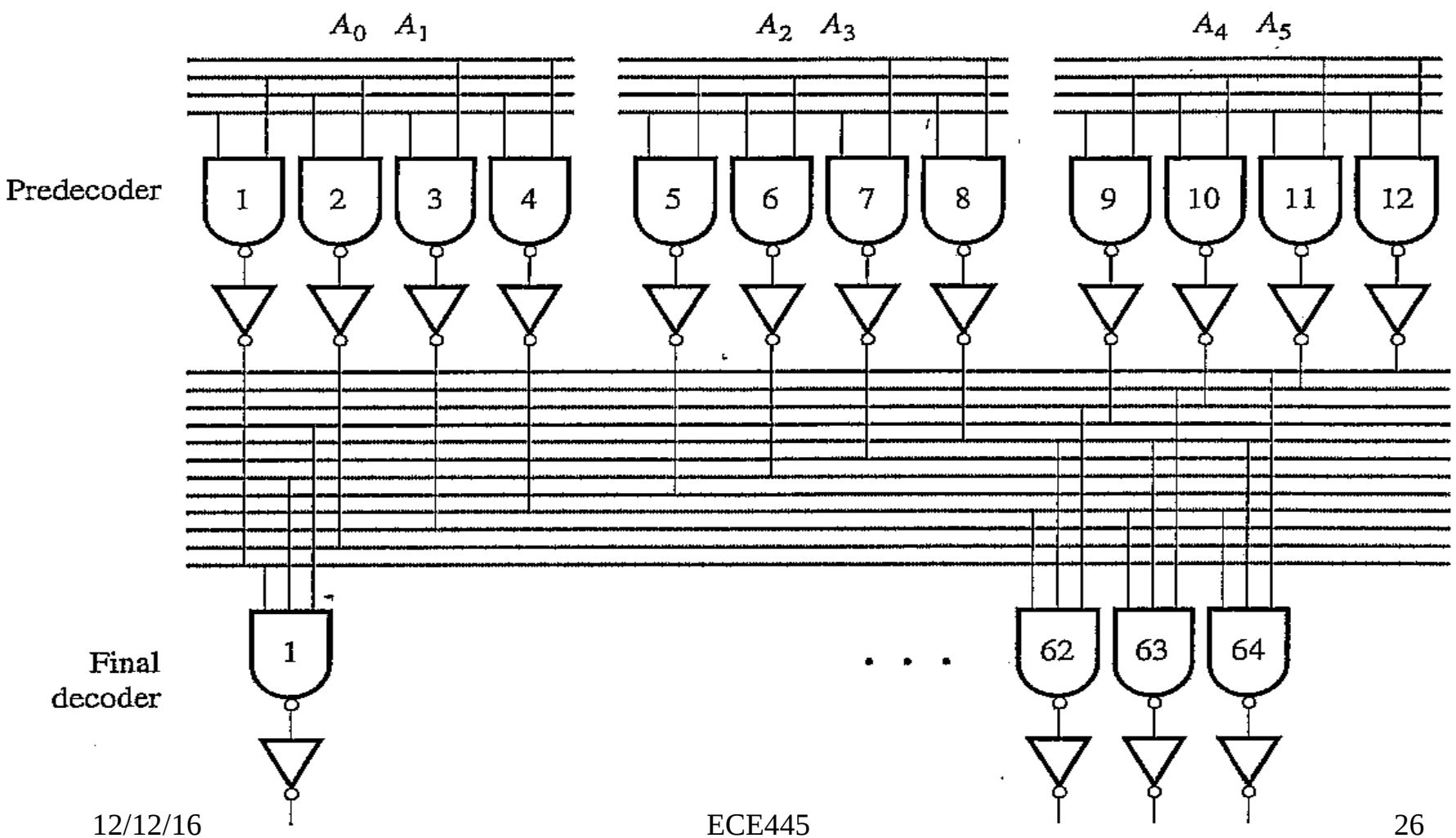
To minimize delay caused by large series resistances → cascade
- predecode stage
- final decode stage

Allows intermediate signals from the predecode stage to be reused by the final decoding stage

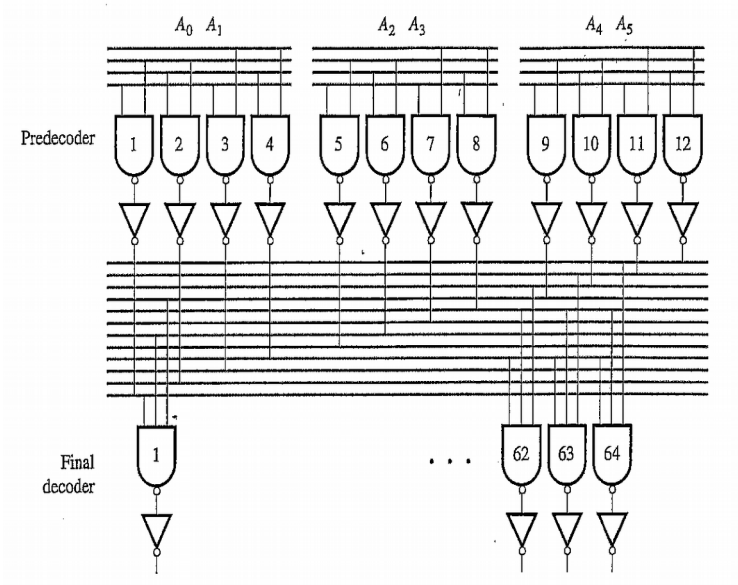
Alternative 6-input NAND



Reuse of Intermediate Signals



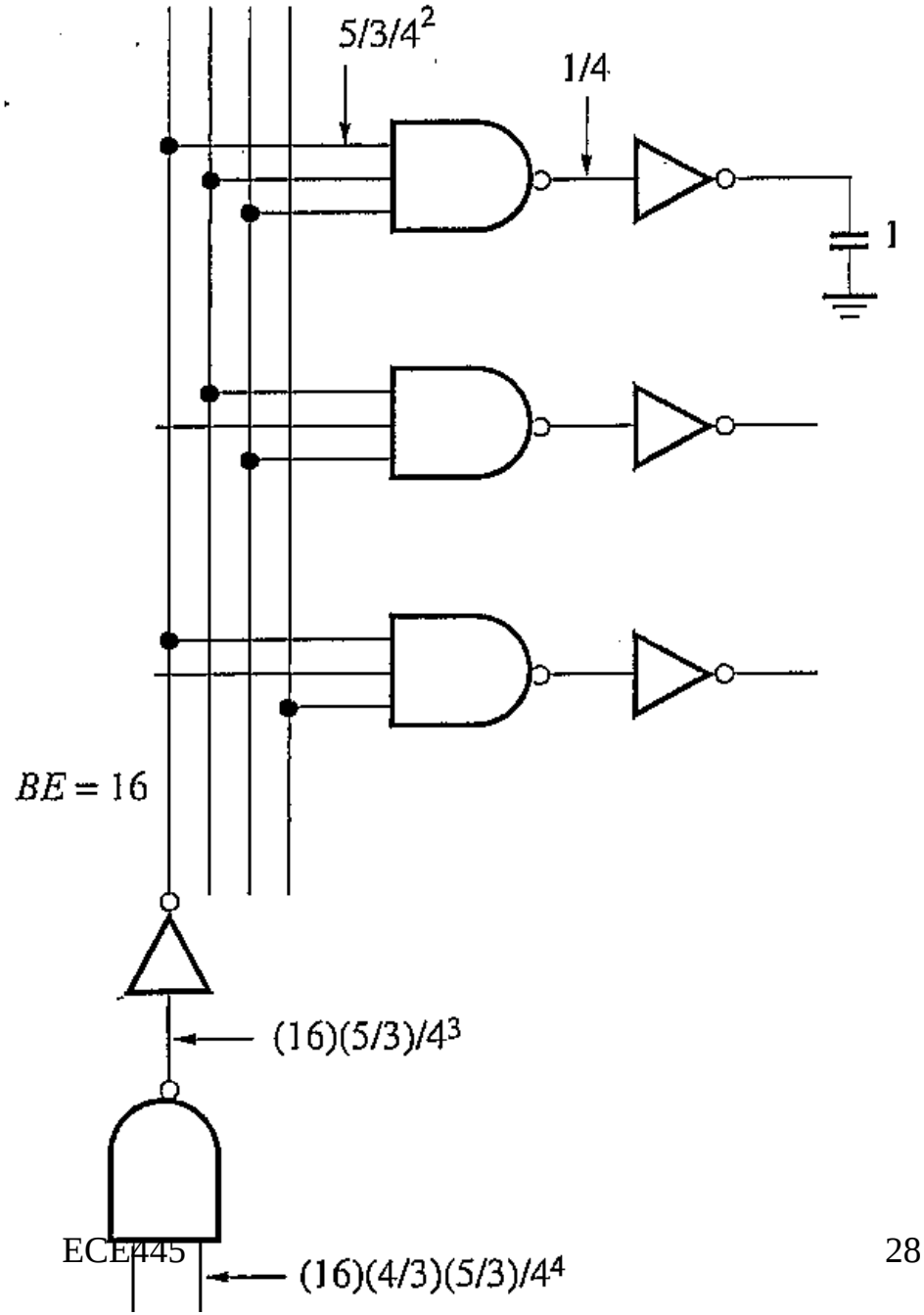
Sizing Using Logical Effort



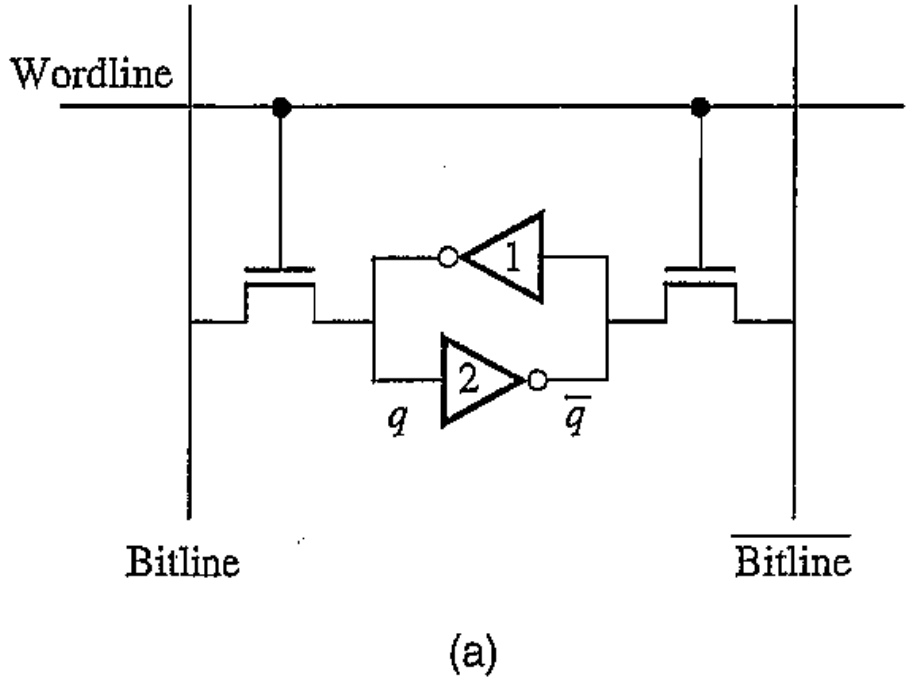
Size the decoder assuming:
- normalized output load = 1
- fan-out = 4

Work backwards from the output to the input:
- Normalized output has a size =
- Inverter input capacitance =
- The Logical Effort of the 3-input NAND =
- Input capacitance of 3-input NAND =
- Branching effort at output of predecoder =
- Input capacitance of the inverter =
- The 2-input NAND has an input capacitance =

Decoder Size by Logical Effort



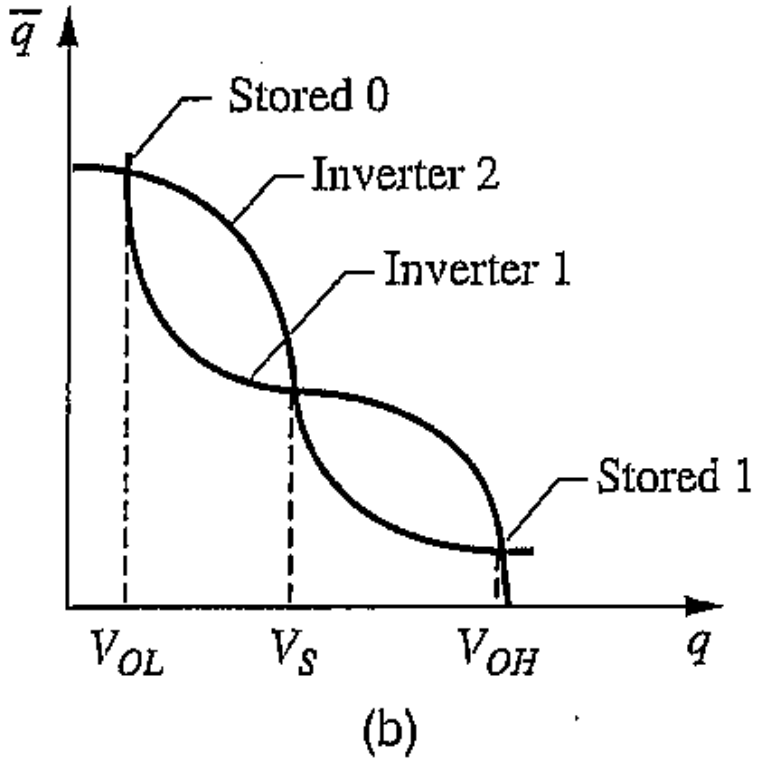
SRAM (Static RAM) Cell



(a)

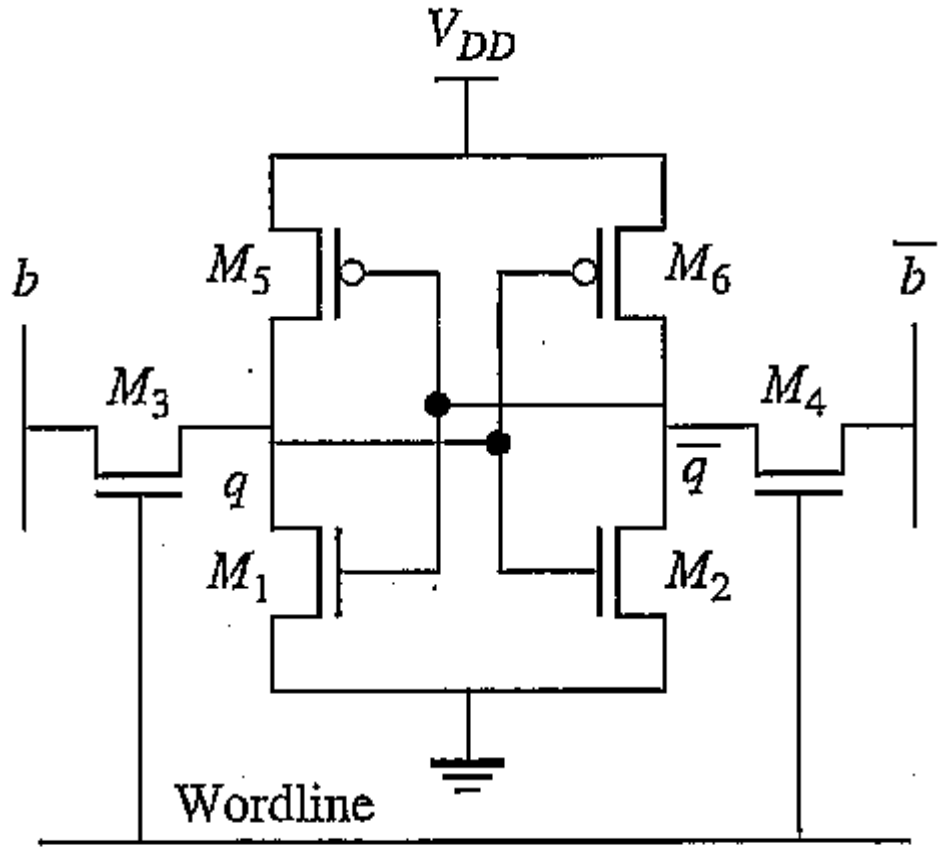
Basic SRAM Cell

Voltage Transfer Characteristics



(b)

6-T Configuration



Cell Design Considerations

1. Minimize Cell Area ←→ Increase cell density
   - replace load resistors (M5 and M6) with transistors
   - utilize TFTs for M5 and M6
   - reduce wasted area

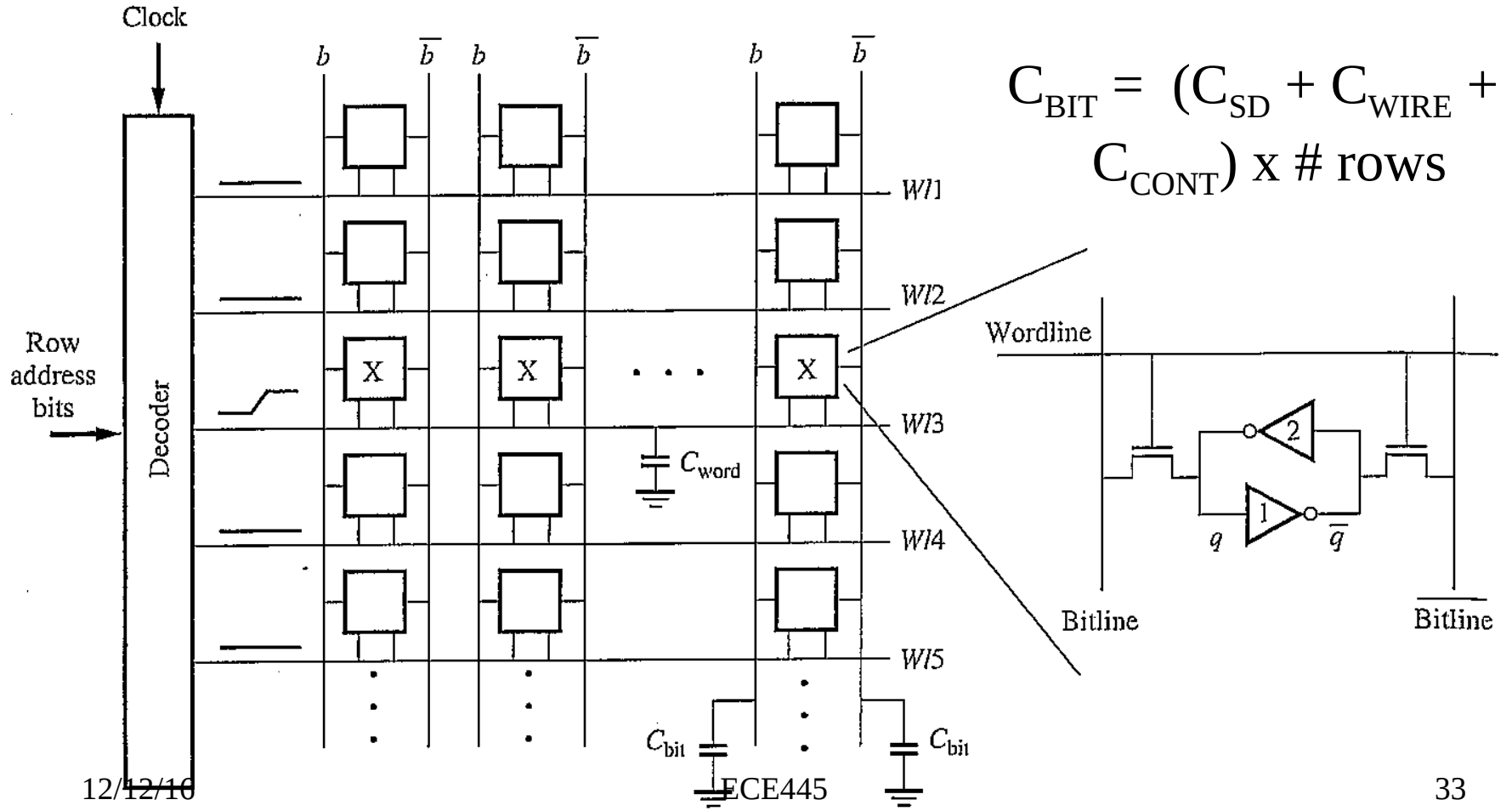
2. Minimize static power → Increase Vt and sub-Vt slope

3. Differential bitlines → improved noise immunity

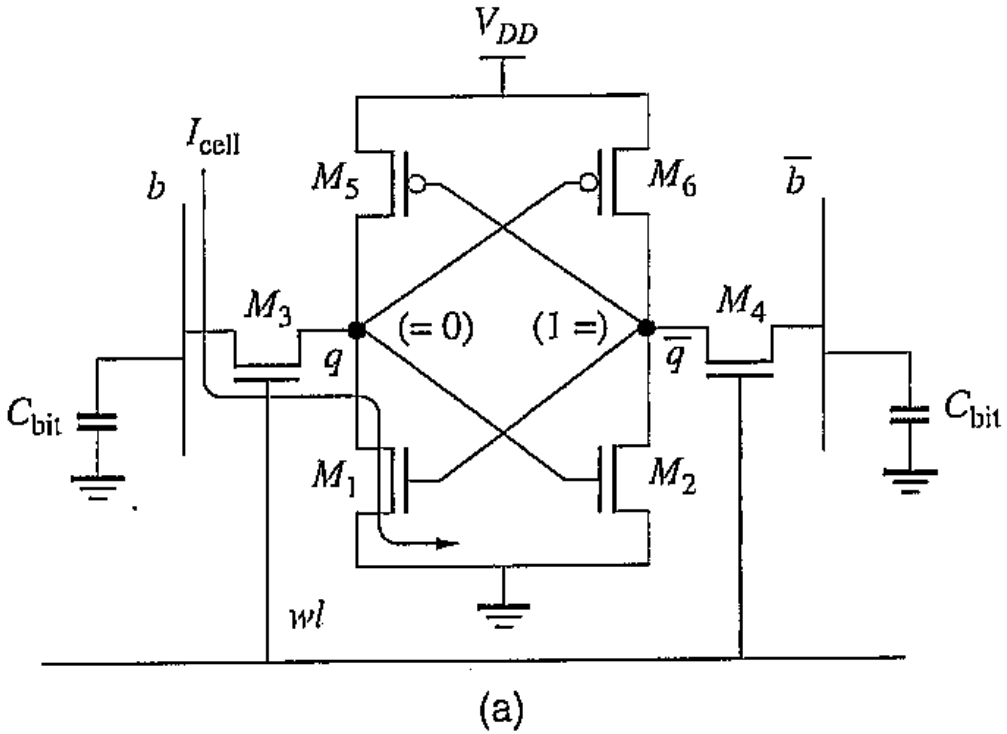
SRAM Array (simplified)

$C_{WORD} = 2 (C_G + C_{WIRE})$ x # columns

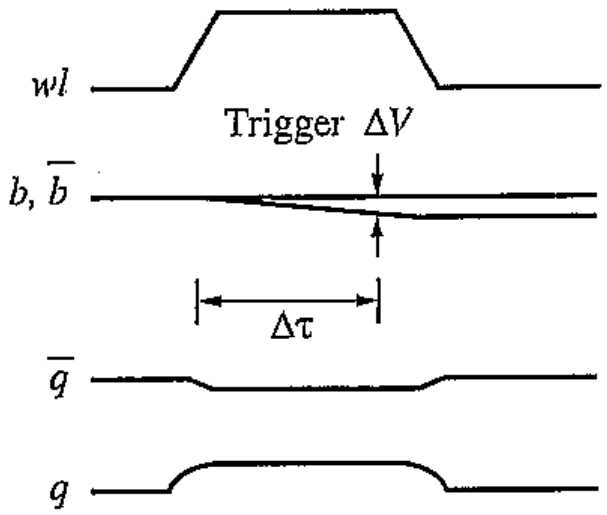
$C_{BIT} = (C_{SD} + C_{WIRE} + C_{CONT})$ x # rows



SRAM Read Operation



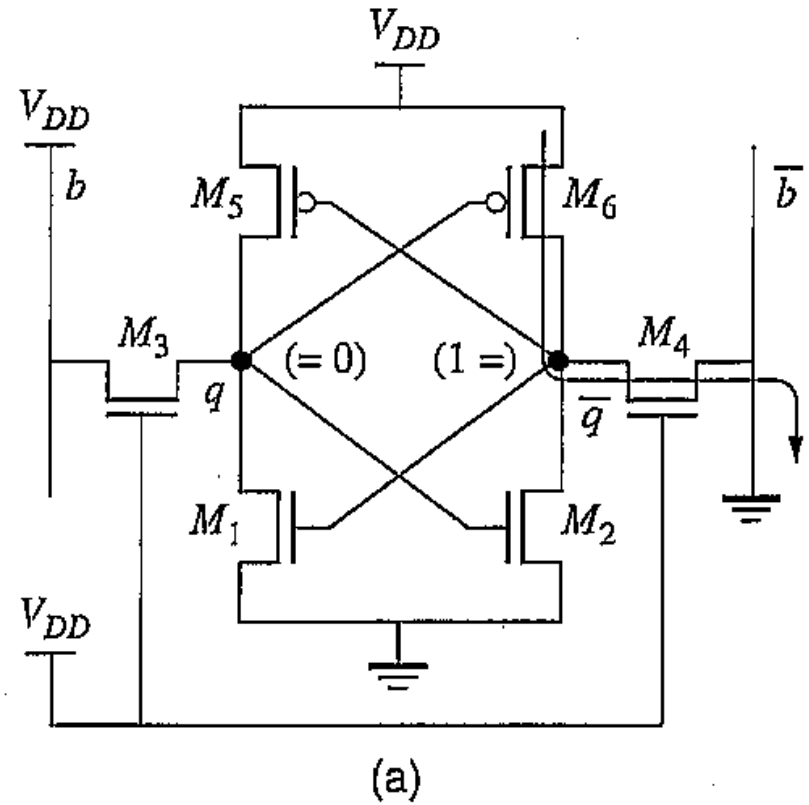
(a)



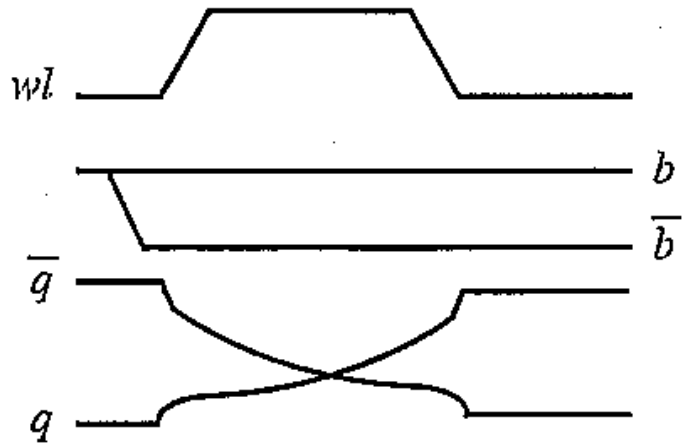
(b)

1. Precharge b, $\overline{b}$ to V_{DD} column pull-up R
2. Turn on wordline to select row
3. CBIT slowly discharges on either b or $\overline{b}$
4. b and $\overline{b}$ → sense amplifier

SRAM Write Operation



(a)



(b)

1. Force either b or $\overline{b}$ low
2. Turn on appropriate word line
3. Node q or $\overline{q}$ is driven low causing the state to switch

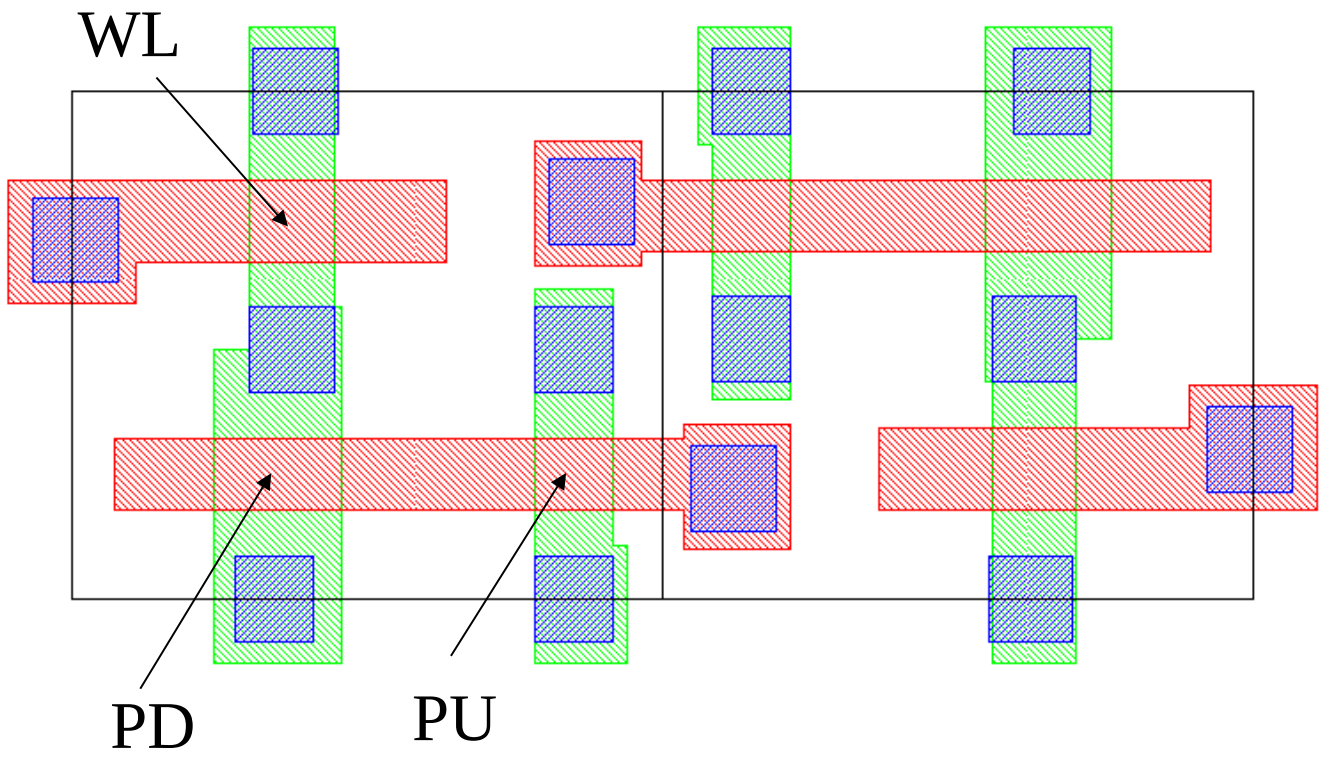
Transistor Sizing Guidelines

1. During read: need to ensure stored value is not disturbed – voltage can not drop below V_S
2. Make conductance of M1, M2 greater than M3, M4
3. Need to supply enough I_{CELL} to discharge C_{BIT} sufficiently in 20-30% of a clock cycle
4. Can estimate I_{CELL}

$$I_{CELL} = C_{BIT} \Delta V / \tau$$

5. During write: make conductance of M3, M4 greater than M5, M6

Cell Layout



SRAM Column I/O Circuitry

Pre-charge (PC) circuitry

- used prior to a read or write operation to pull bit lines ‘high’
- equalize b and $\overline{b}$ potentials

Address transition detection (ADT) circuitry to generate PC signal

- triggered by any transition of the address inputs

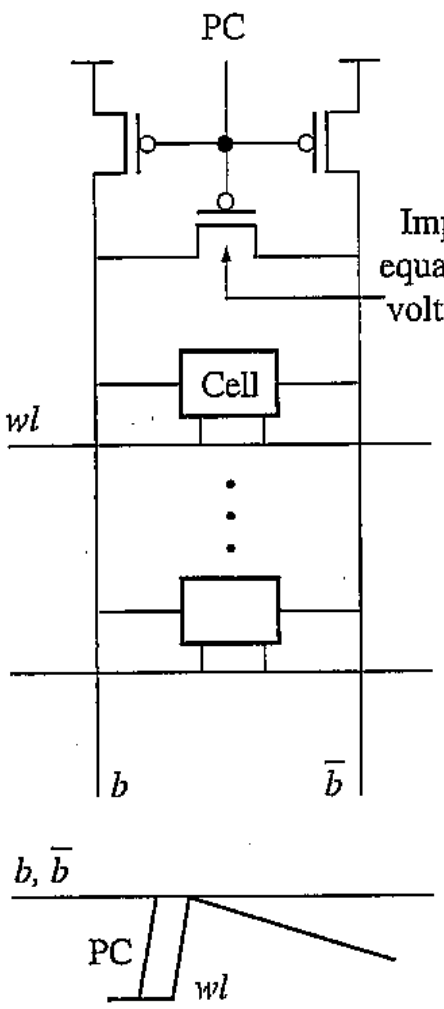
Column decoder – multiplexer

Read and Write circuitry

- R/W selection circuit
- sense amplifiers for read operation
- write drivers for write operation

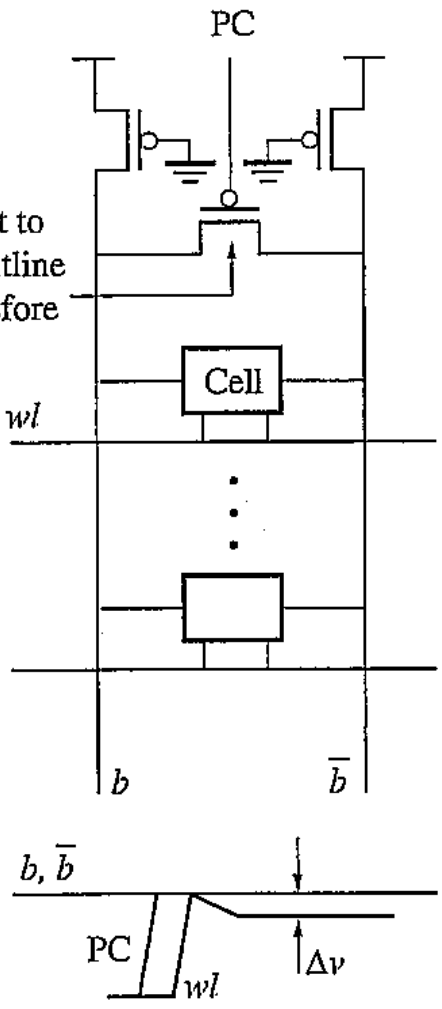
Column Pull-up Circuitry

Dynamic Logic Precharge P



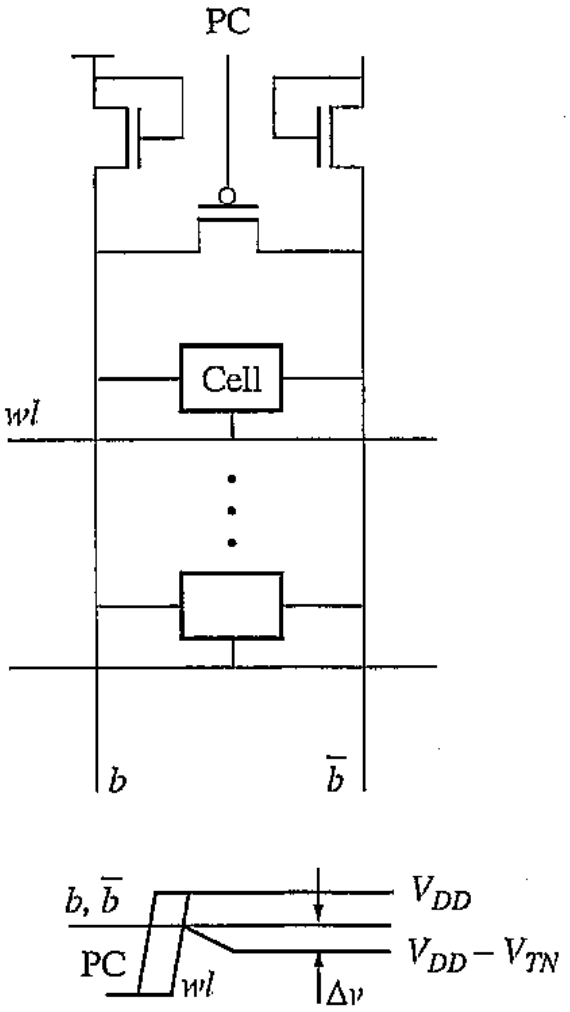
(a)

seudo-NMOS Precharge



(b)

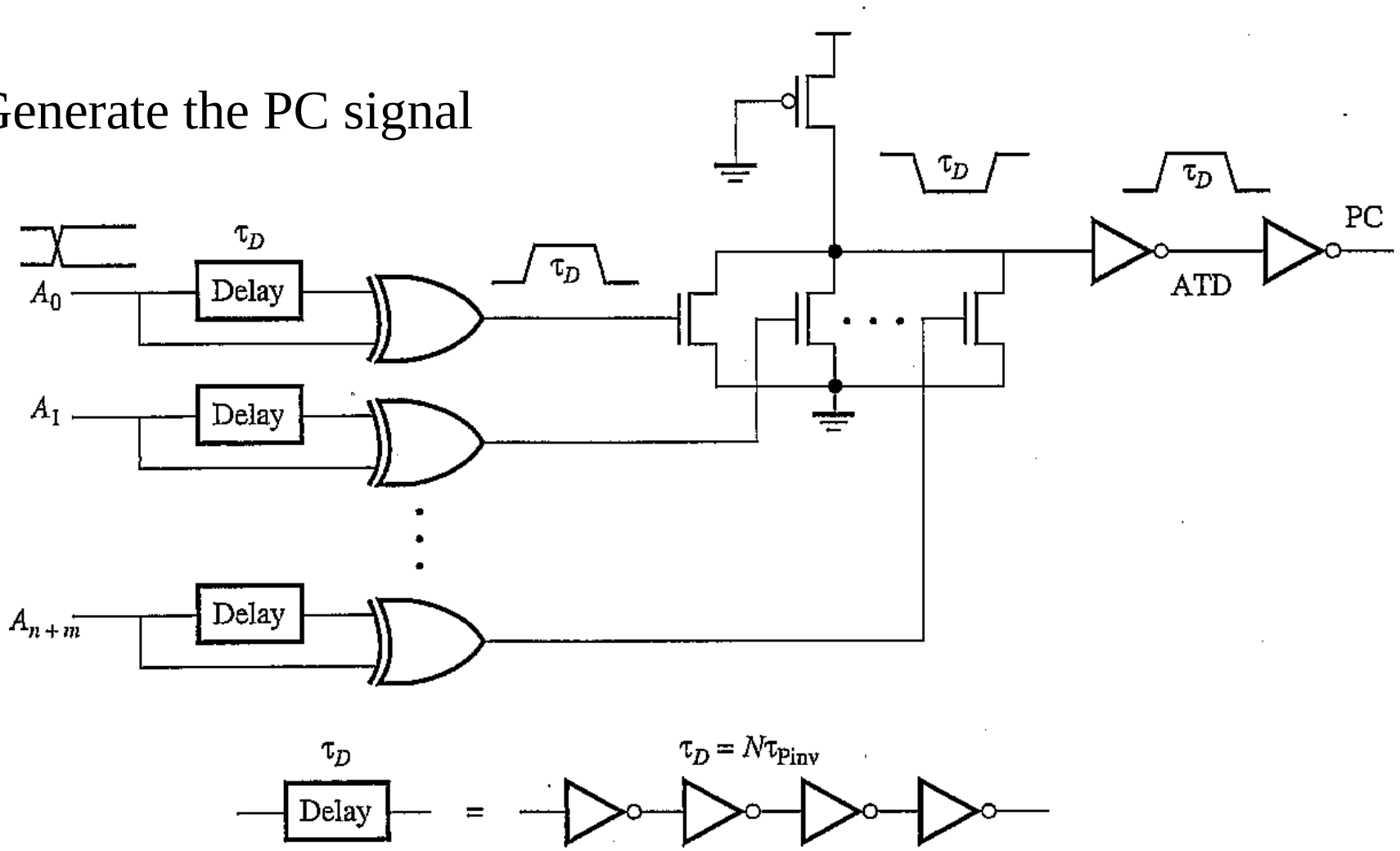
NMOS Saturated Enhancement Load



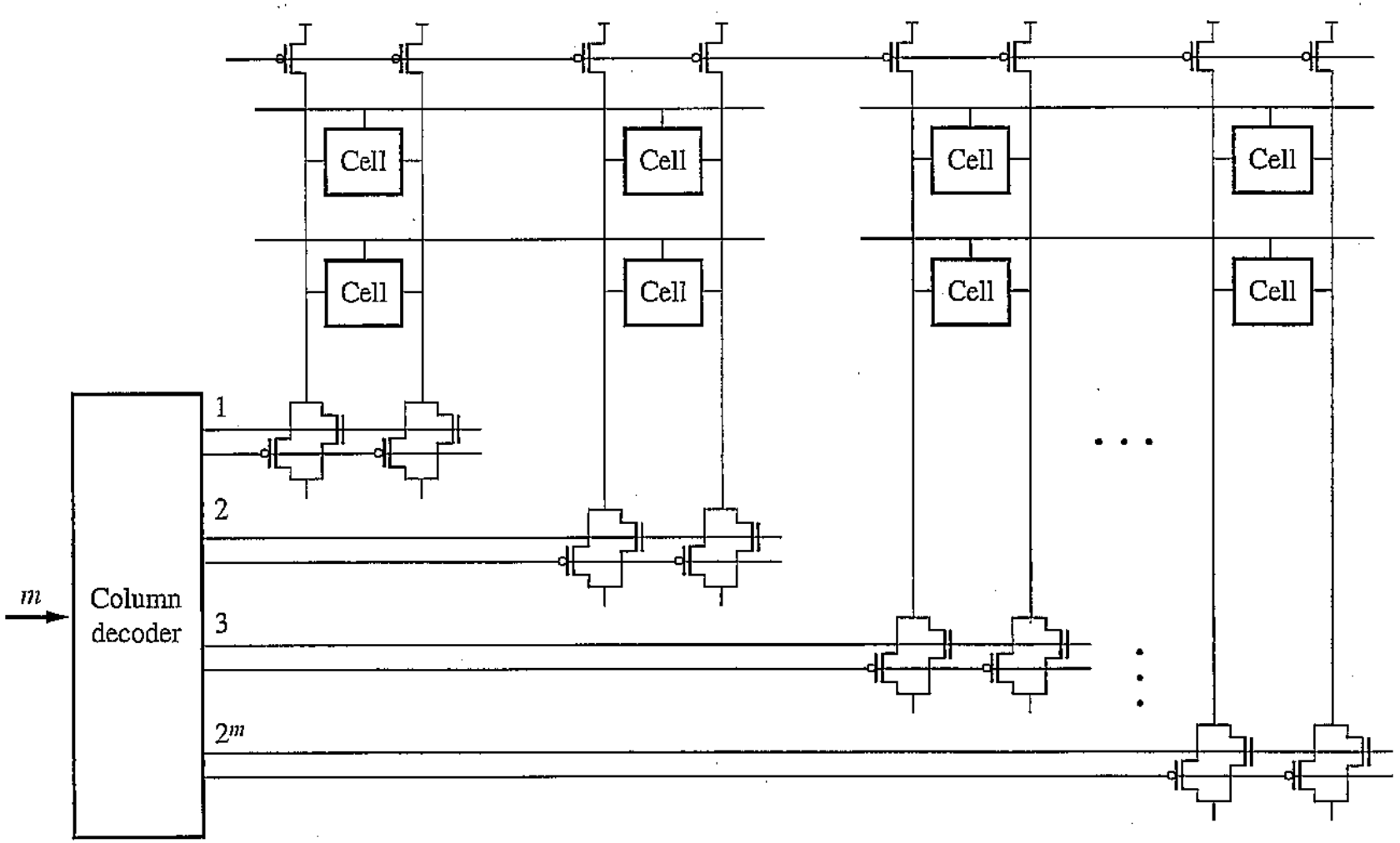
(c)

Address Transition Detection (ADT)

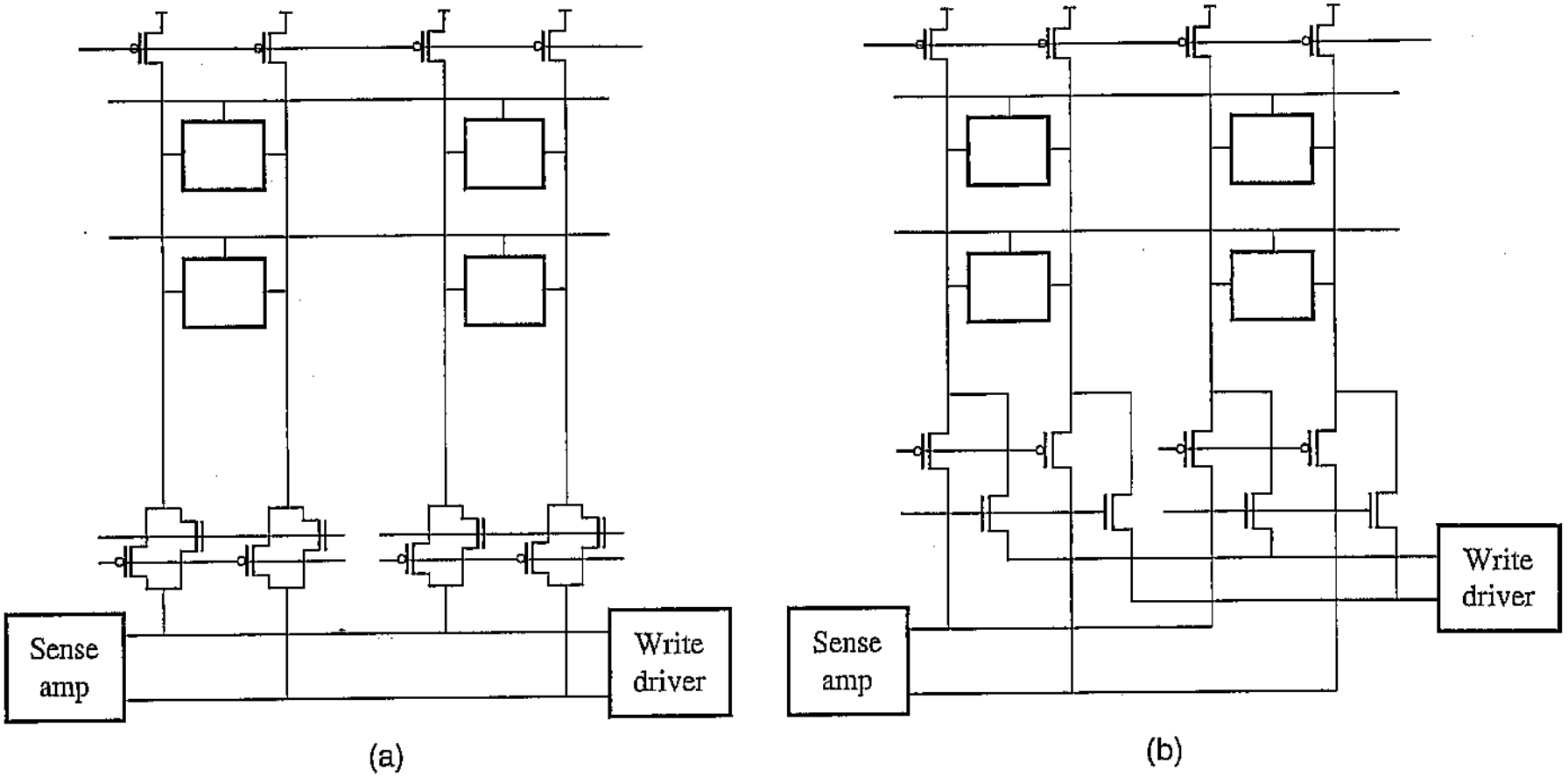
Generate the PC signal



Column Decoder / Multiplexer

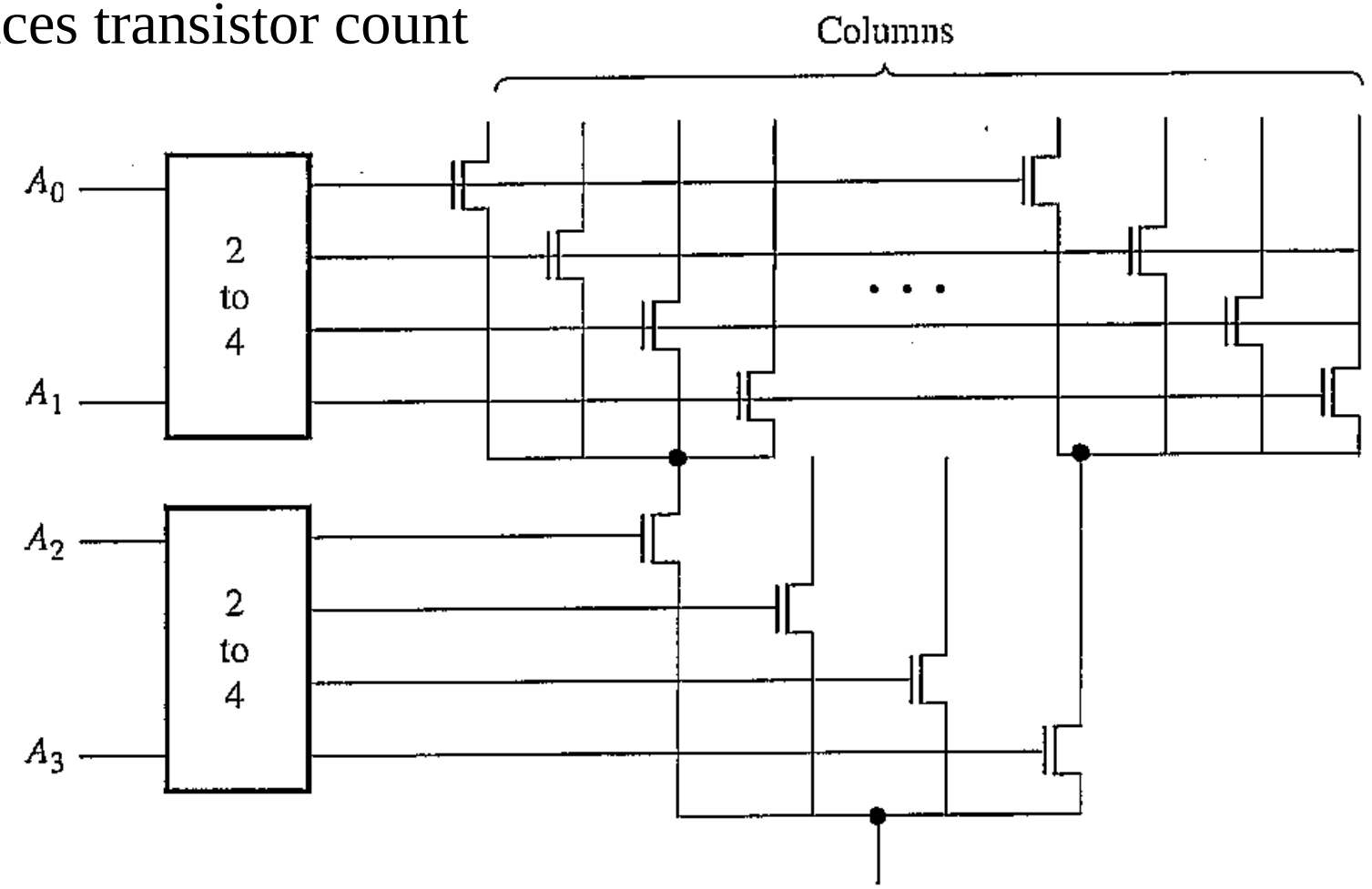


Column Selection / Read - Write

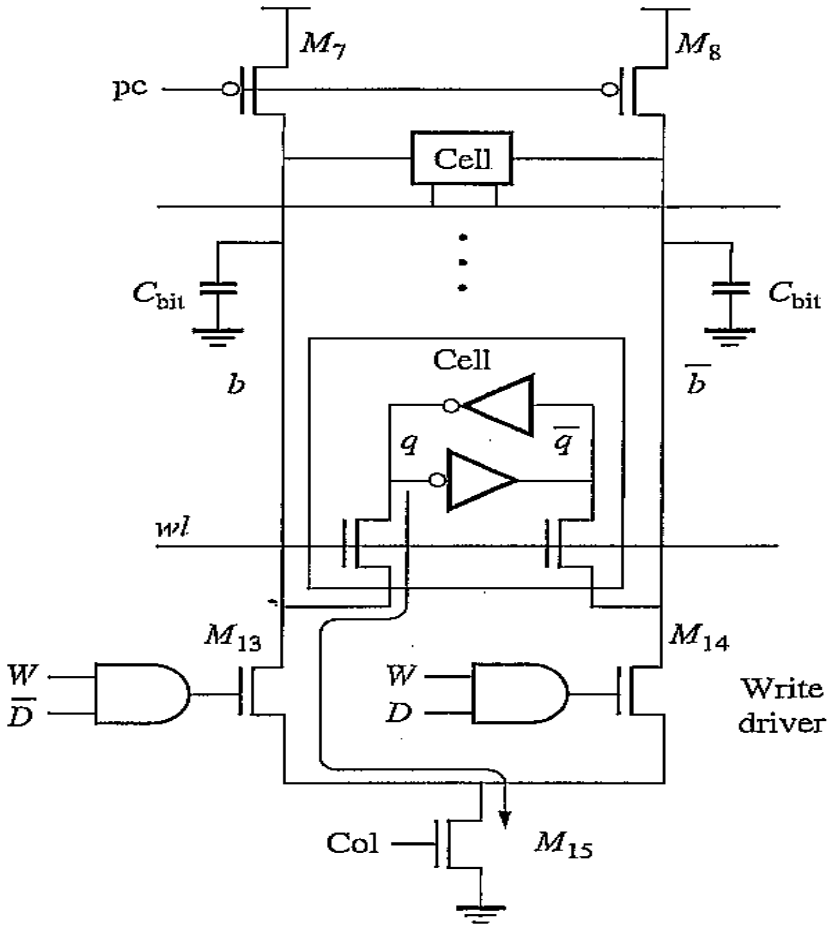


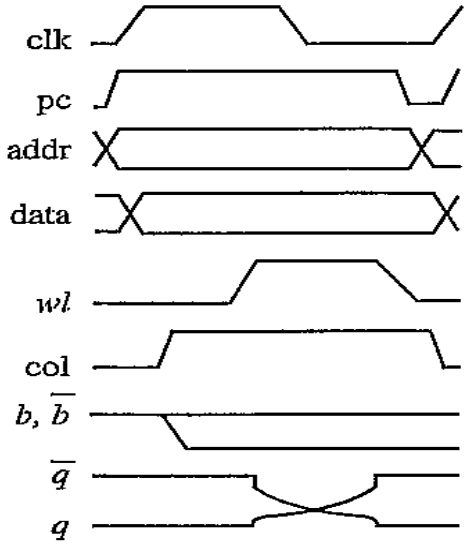
Tree Decoding Structure

Reduces transistor count

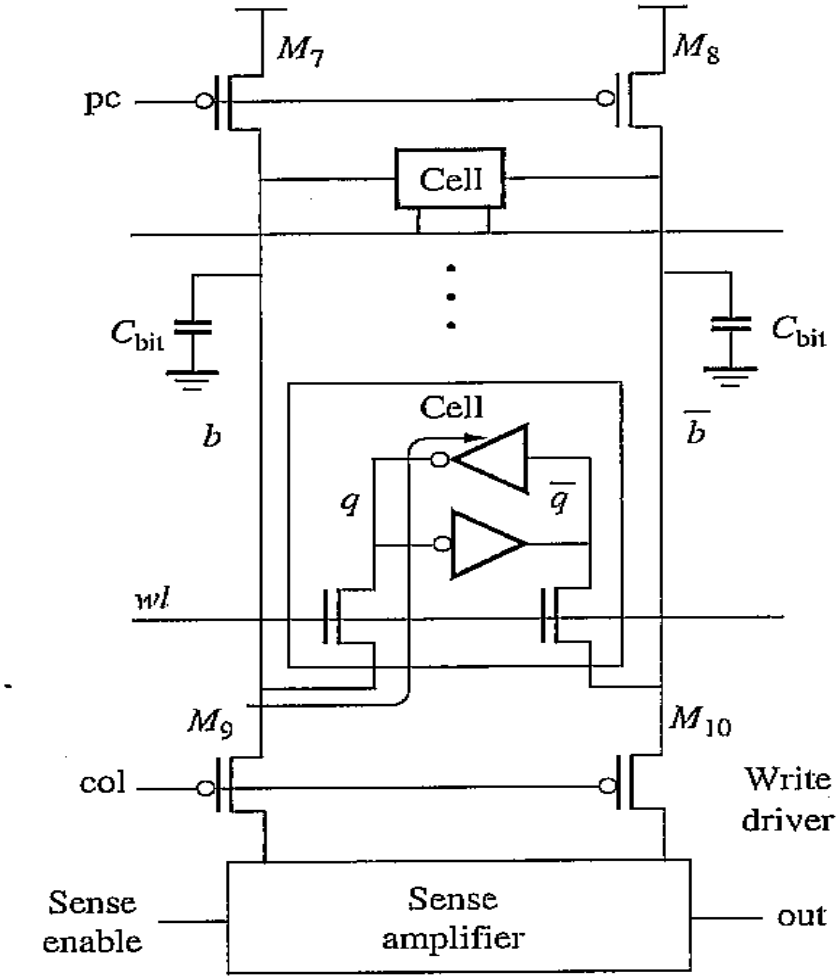


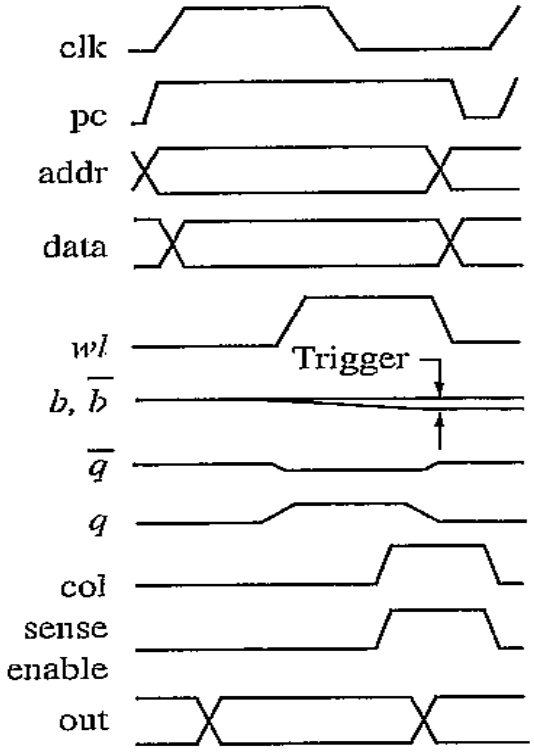
Write Driver Circuitry



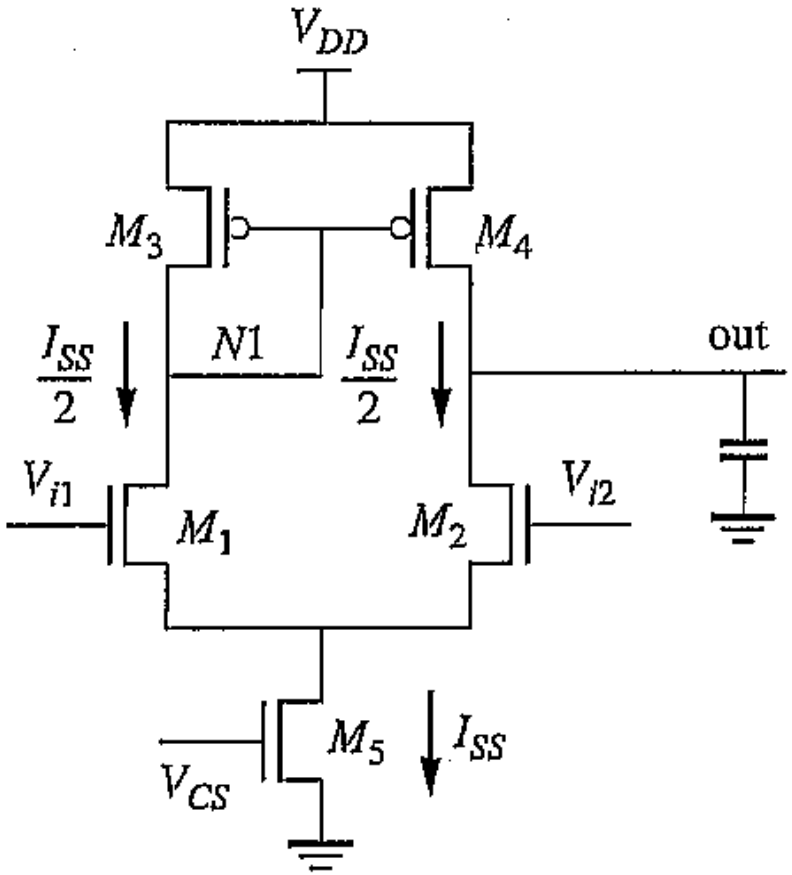


Basic Read Circuitry





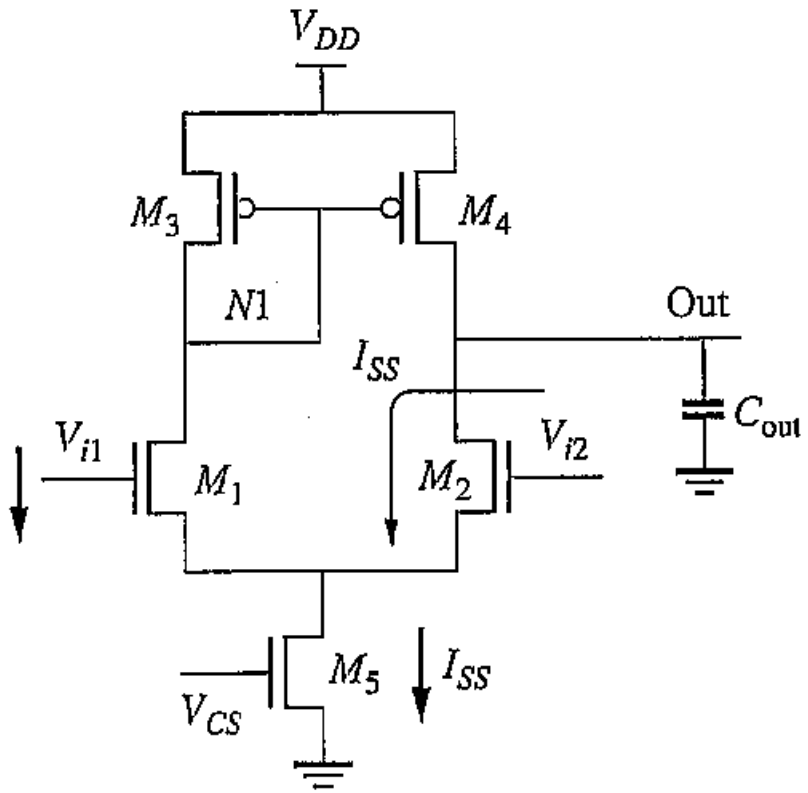
Differential Voltage Sense Amp



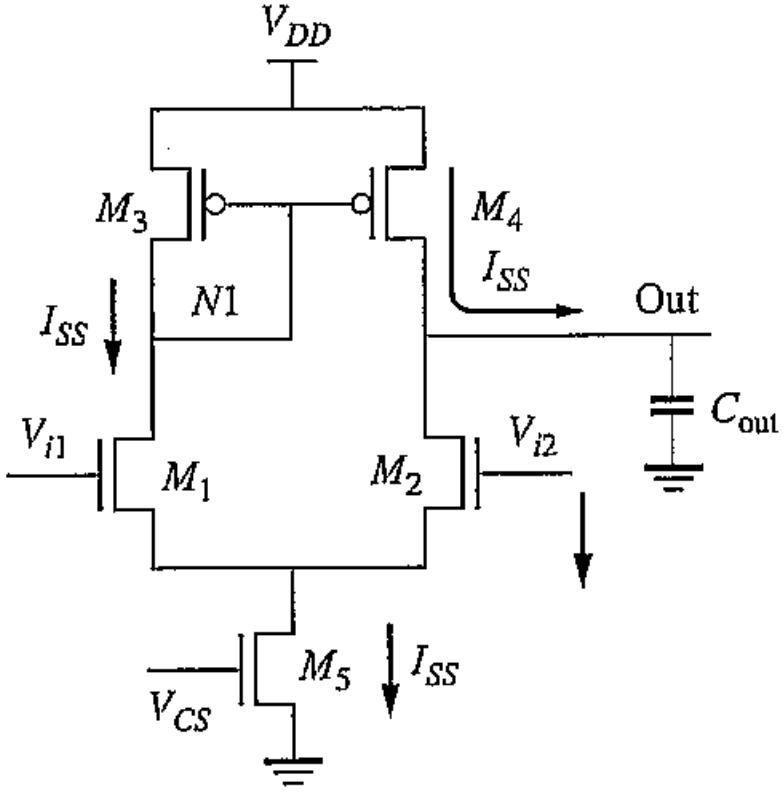
Consists of:
- common source amp
- current mirror
- current biasing

Advantages:
- High Speed Operation
- Good noise immunity
- high CMRR

Detection of ‘0’ and ‘1’

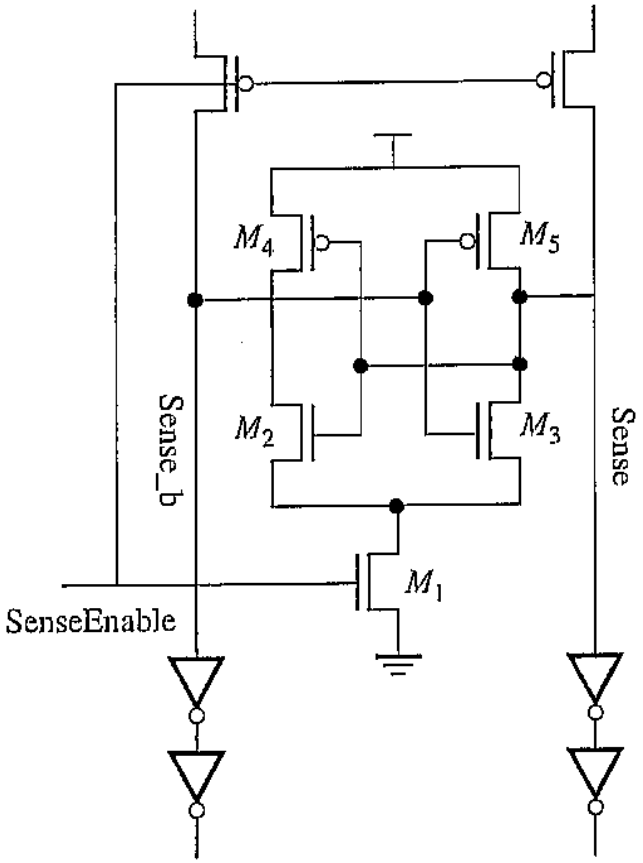


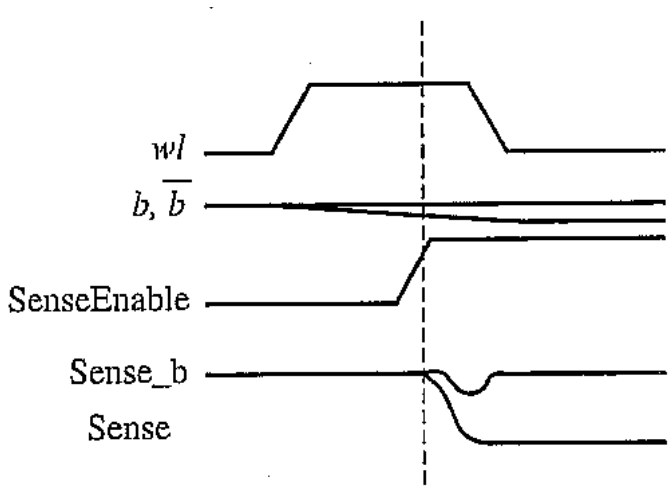
(a) Discharging output



(b) Charging output

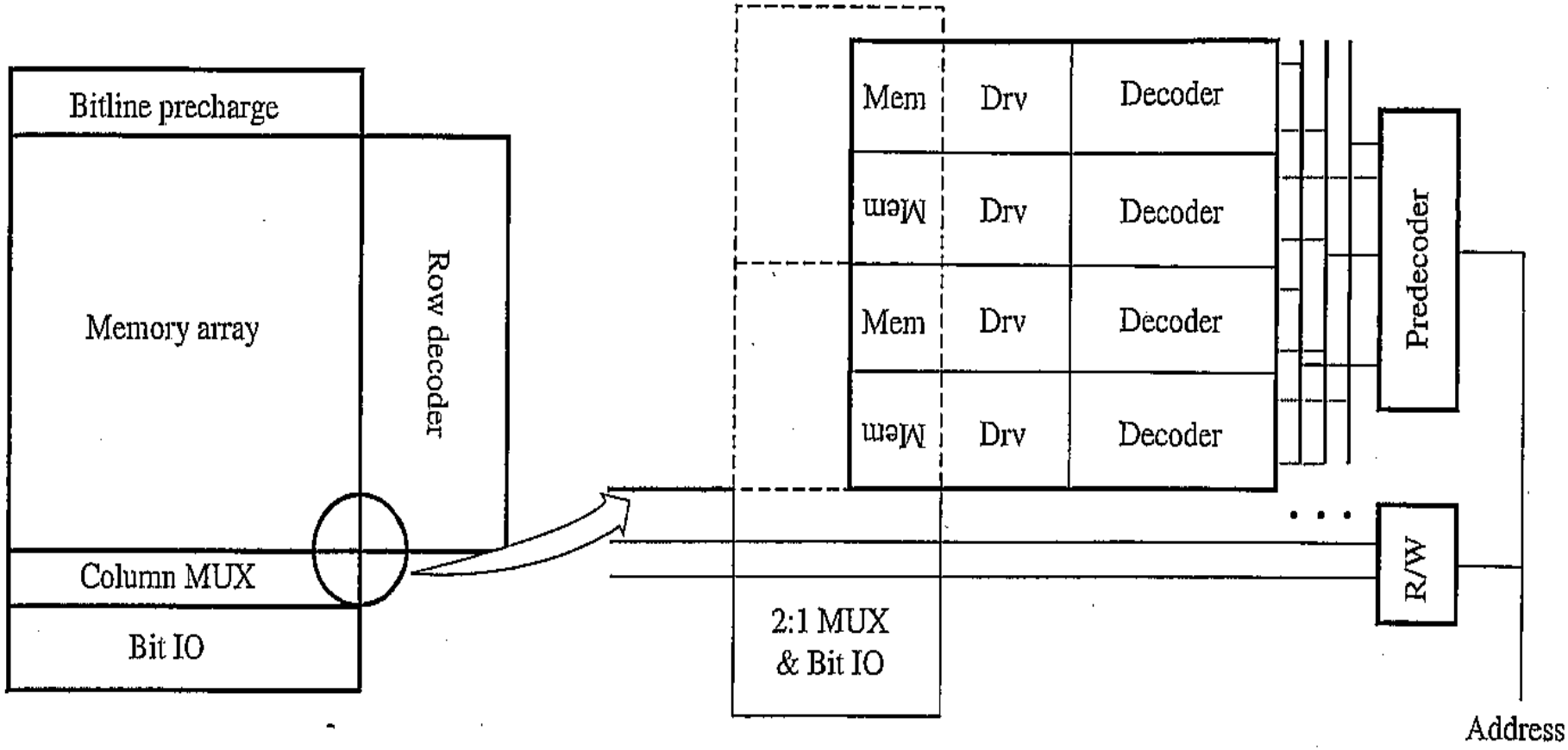
Latch-Based Sense Amp



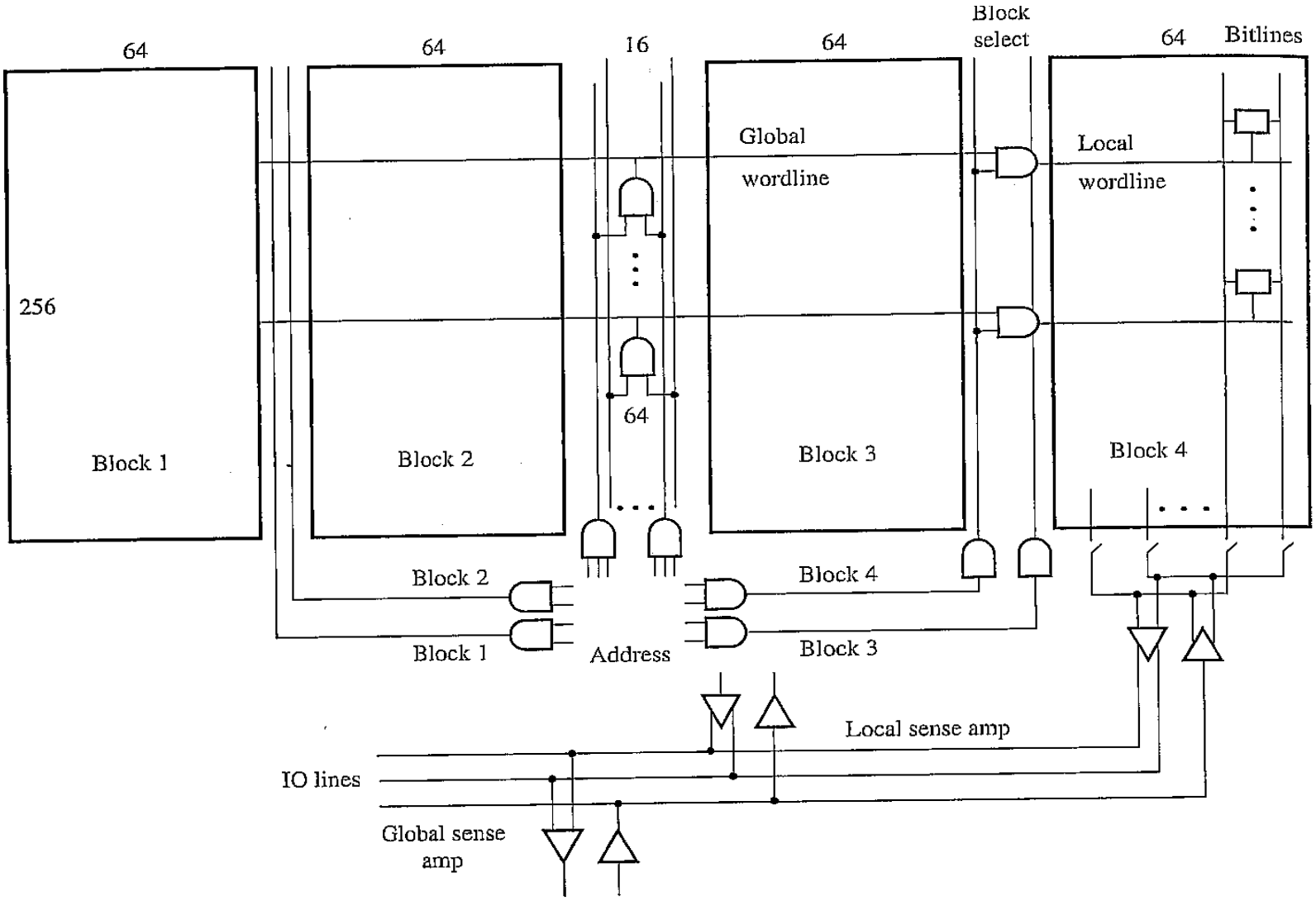


Memory Architecture

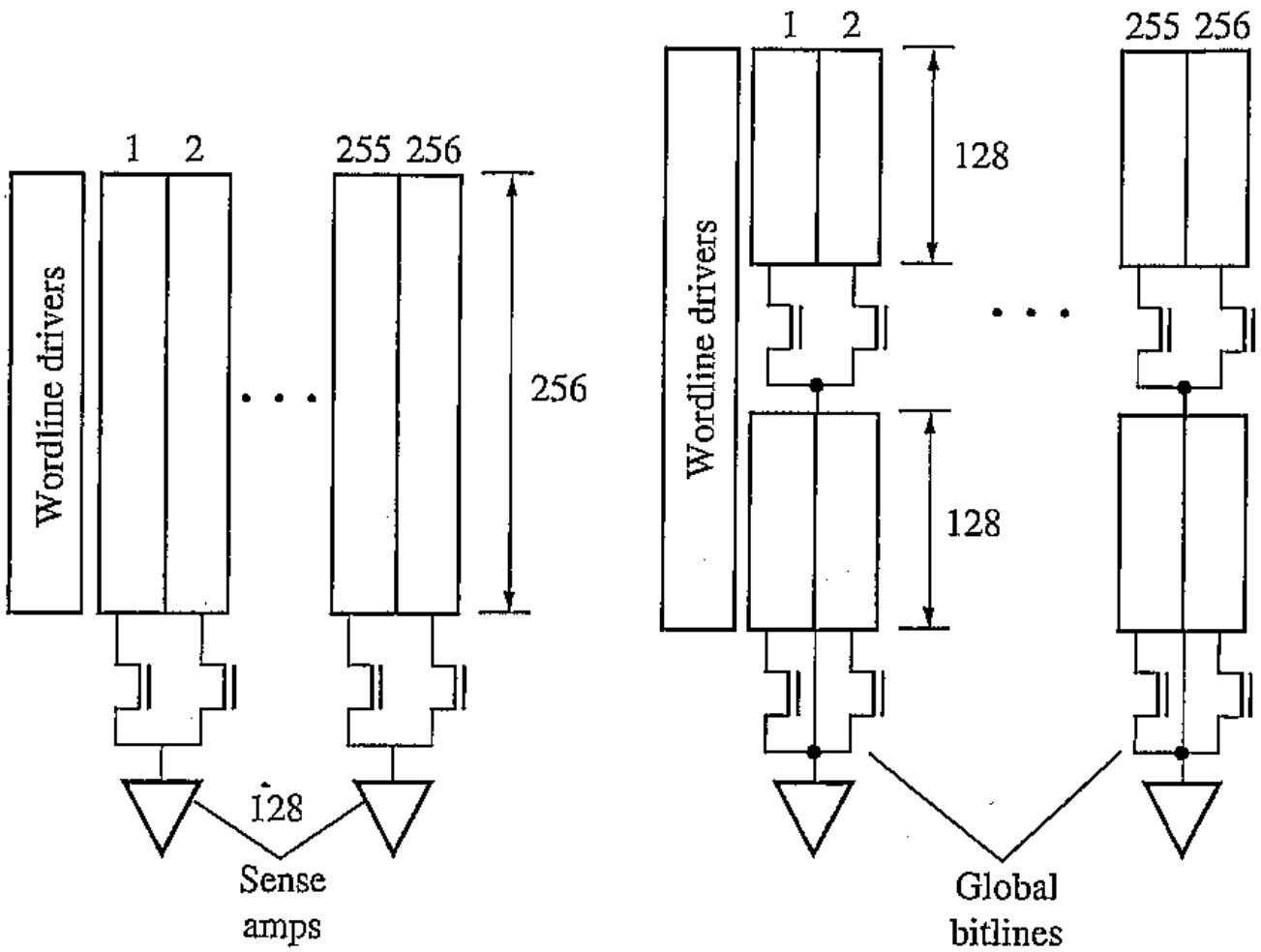
Basic Memory Architecture



Partitioning - Divided Wordline



Partitioning – Divided Bitline



Other Memory Families

Content-Addressable Memories CAM
Dynamic Random Access Memory DRAM
Flash Memory

Associative Memories
Content Addressable Memories (CAMs)

Basic architecture is similar to SRAM

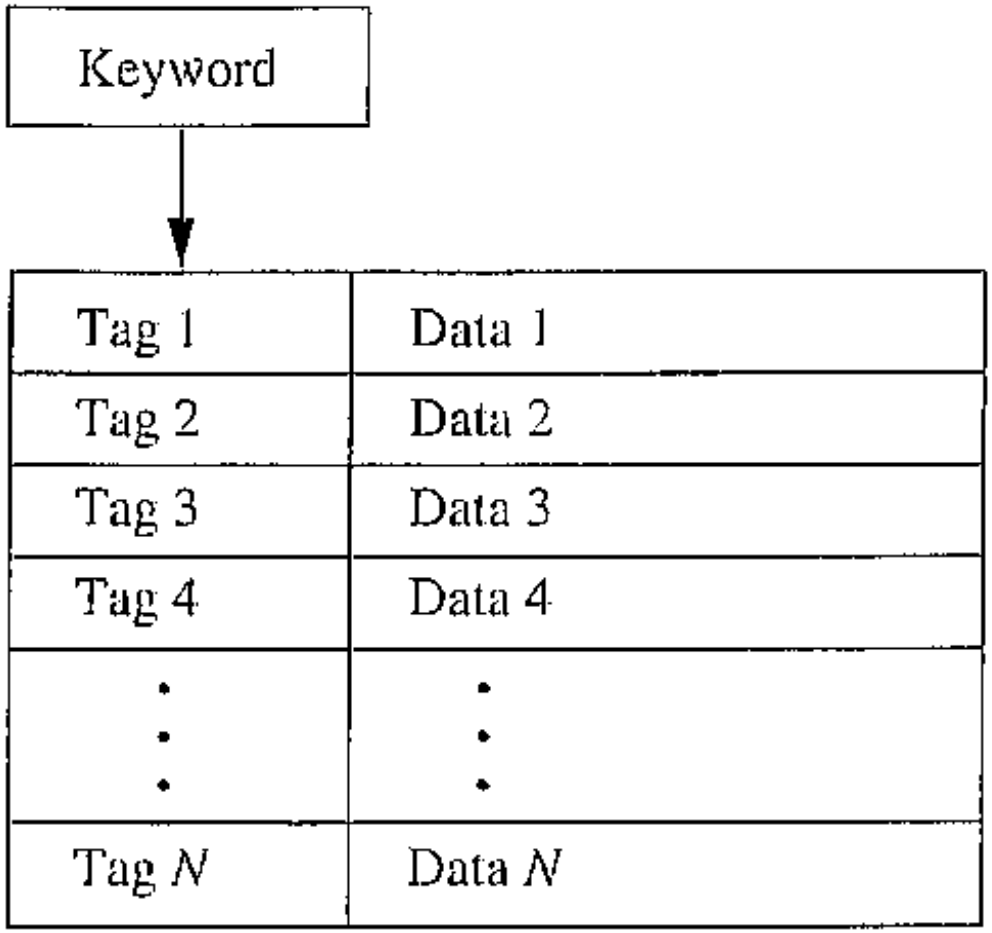
Data we seek is associated with a known binary keyword rather than a known binary address

Compare known keyword against previously stored keywords, known as tags, that reference the data we seek

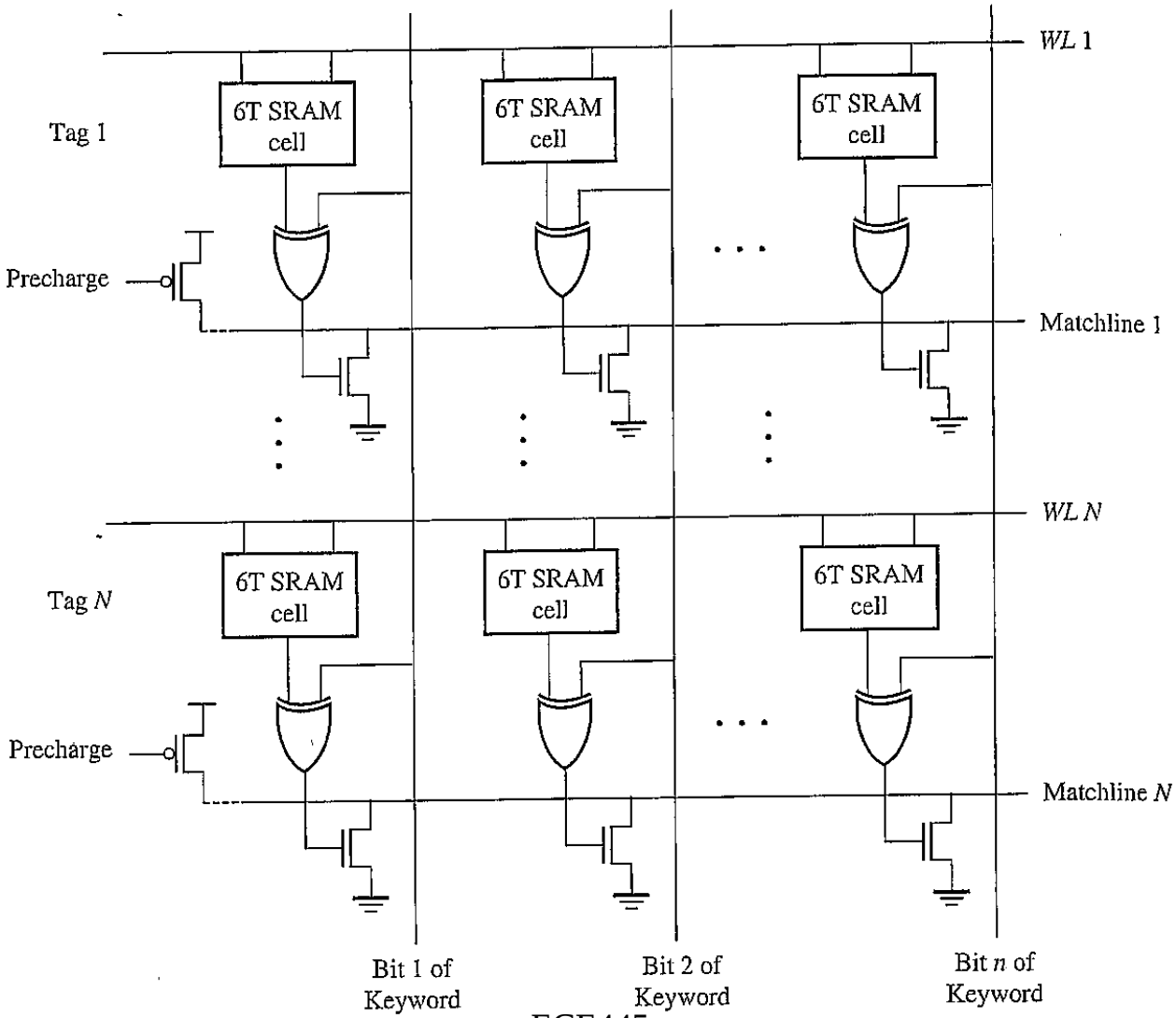
Address decoder → matching scheme

Often used in:
- *cache* memories
- Internet routing tables

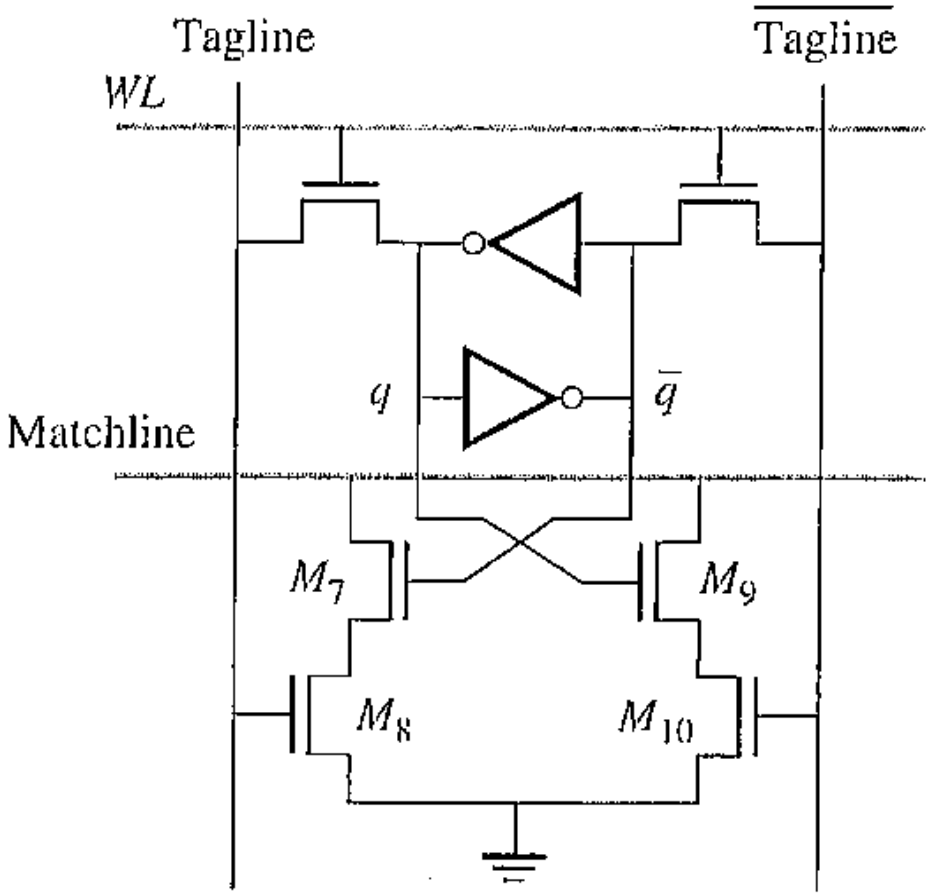
Concept of a CAM Array



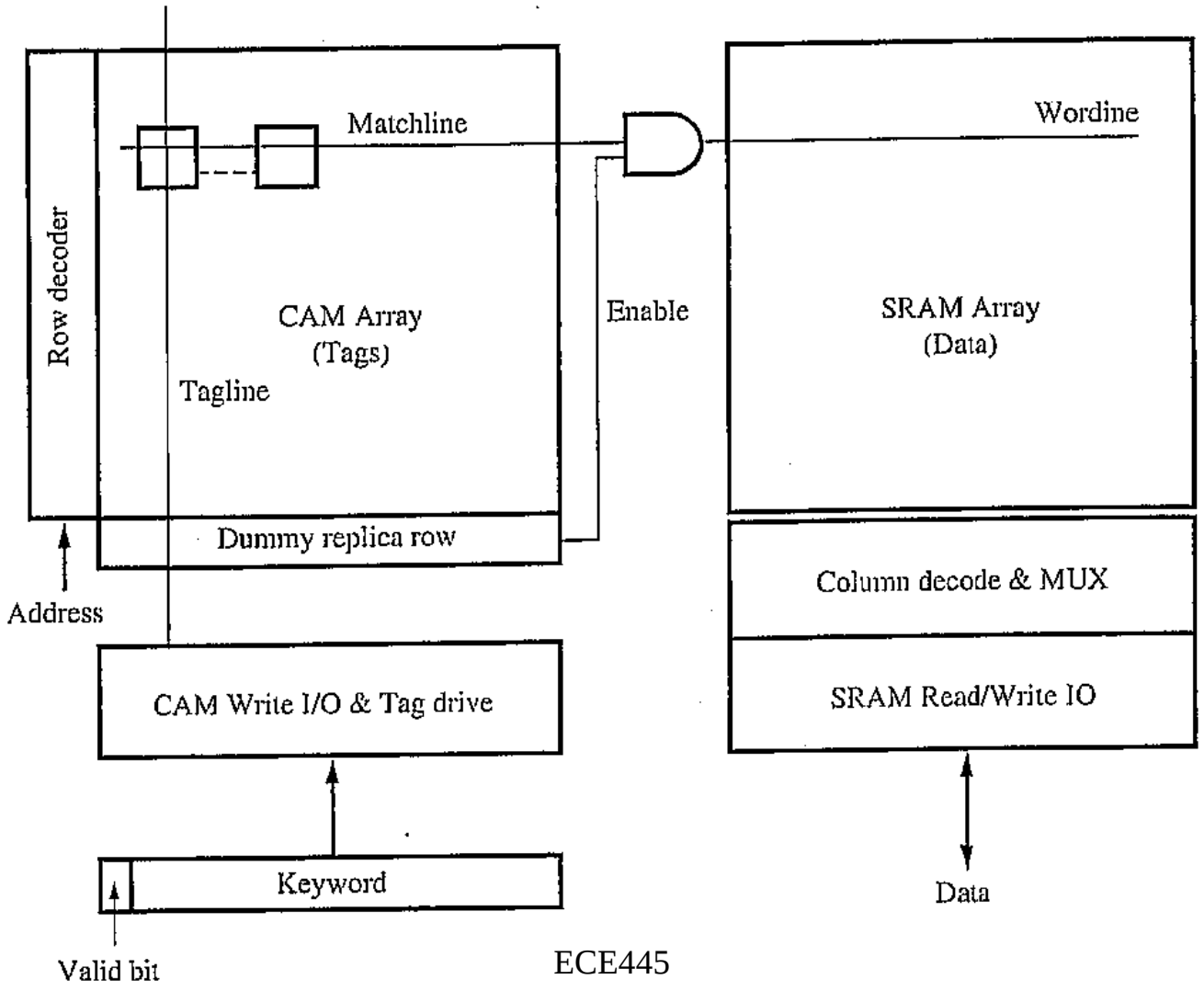
CAM Lookup Array



CAM Memory Cell

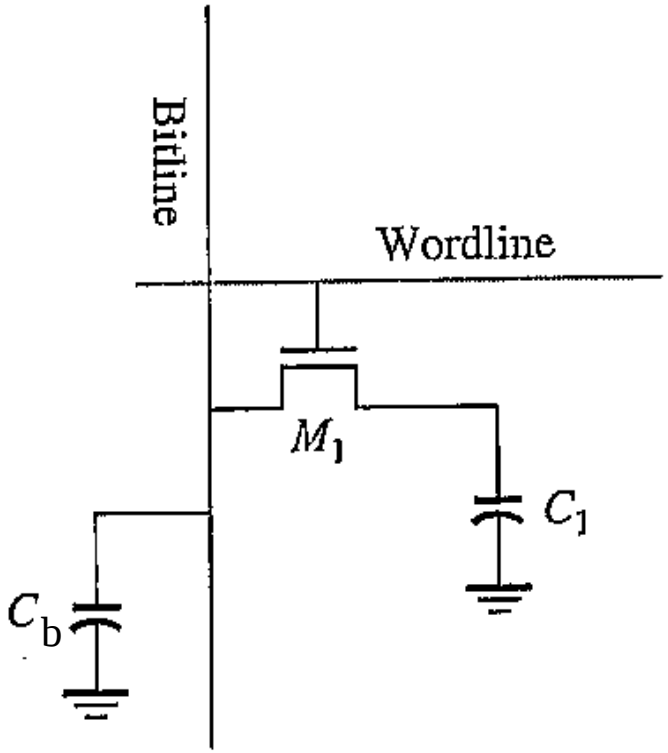


CAM Implementation

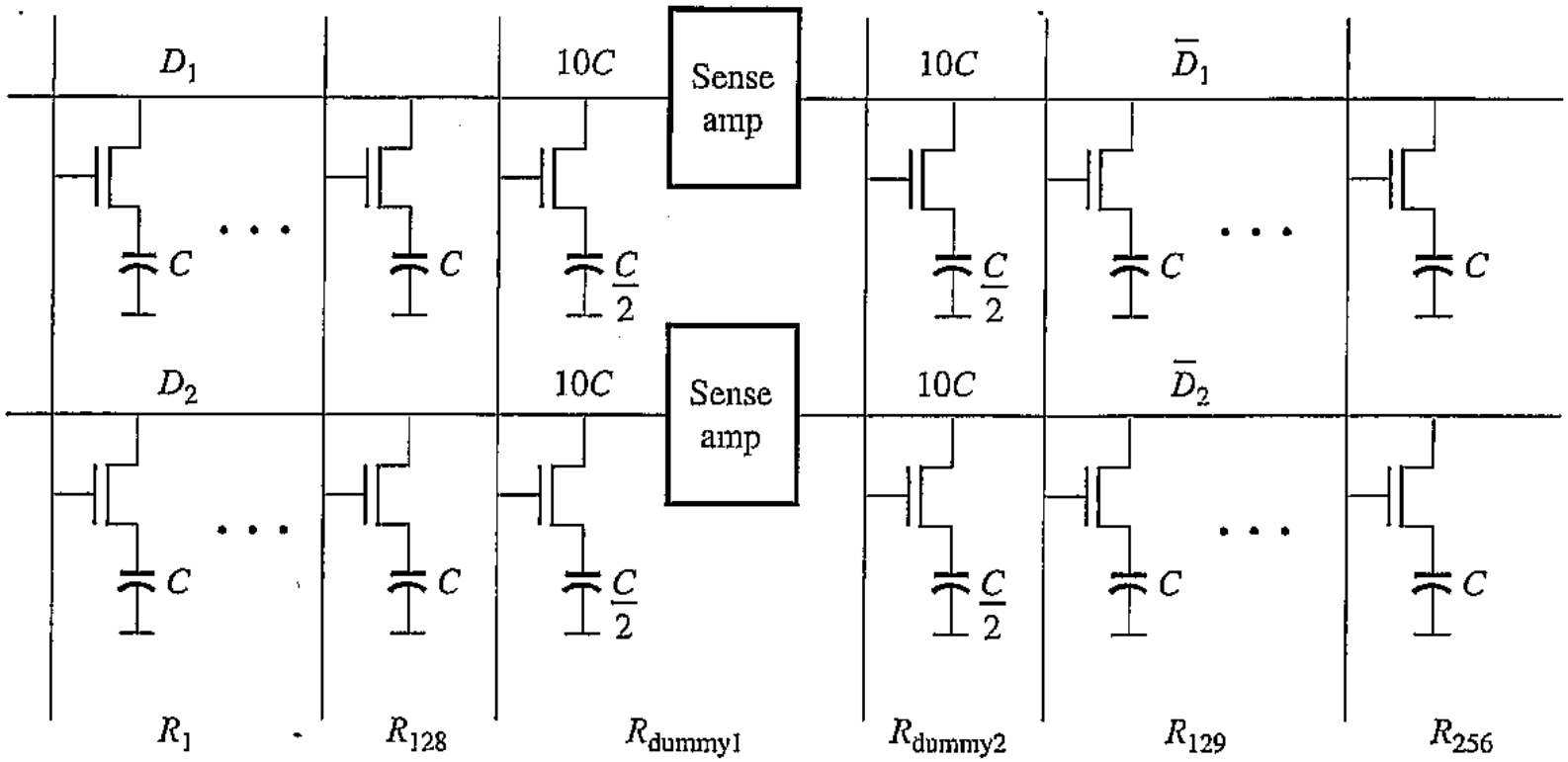


DRAM

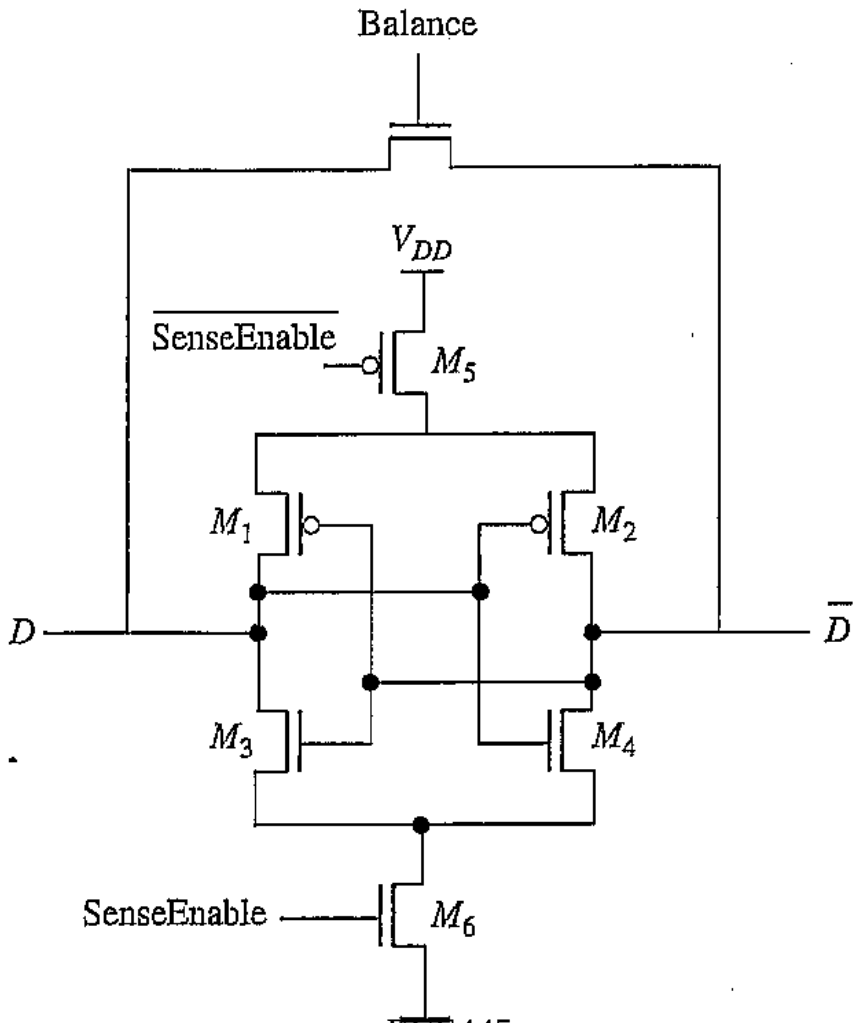
1T DRAM Cell



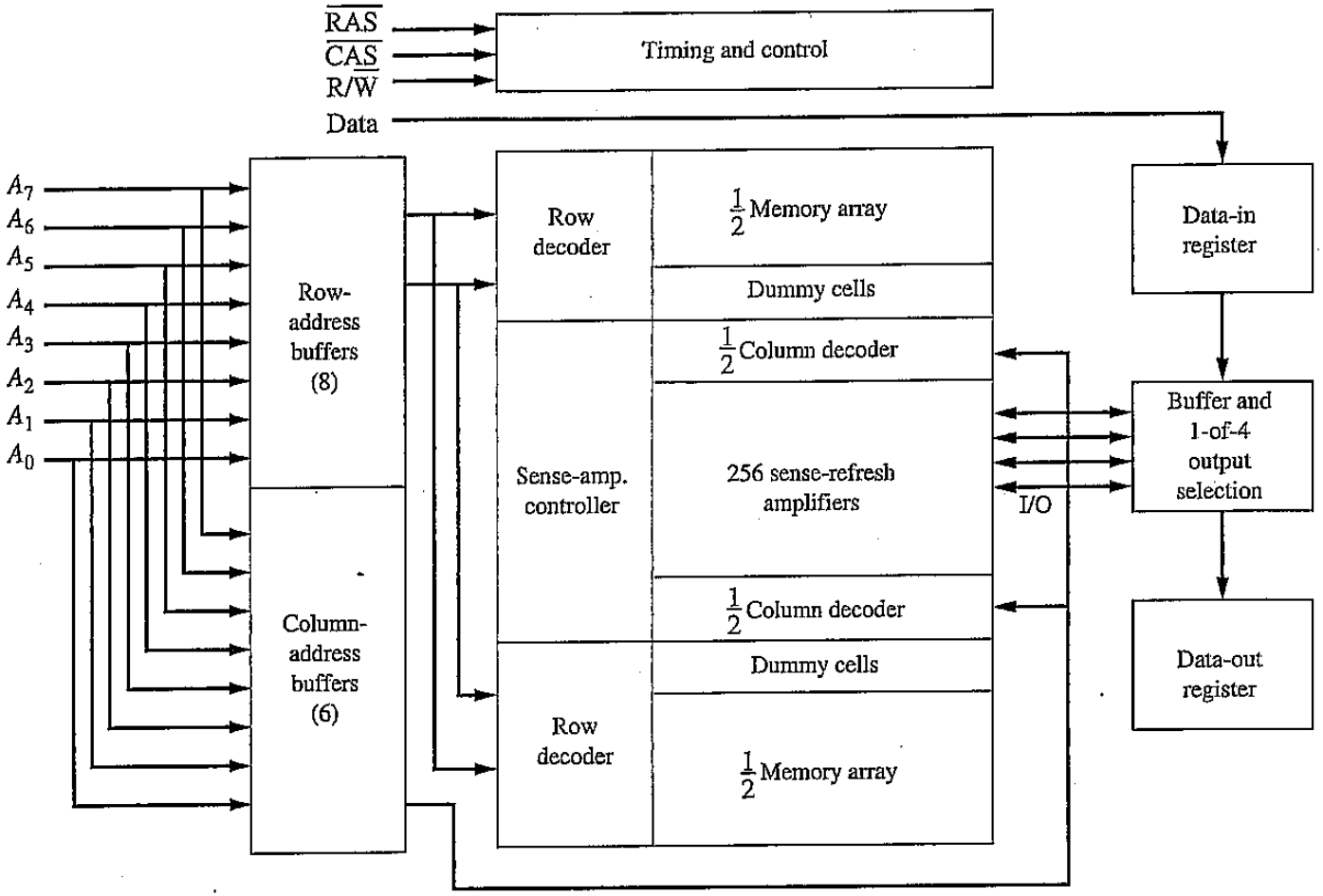
DRAM Array Configuration



DRAM Sense/Refresh Circuit



64-kbit DRAM – Block Diagram



Flash Memory

Applications

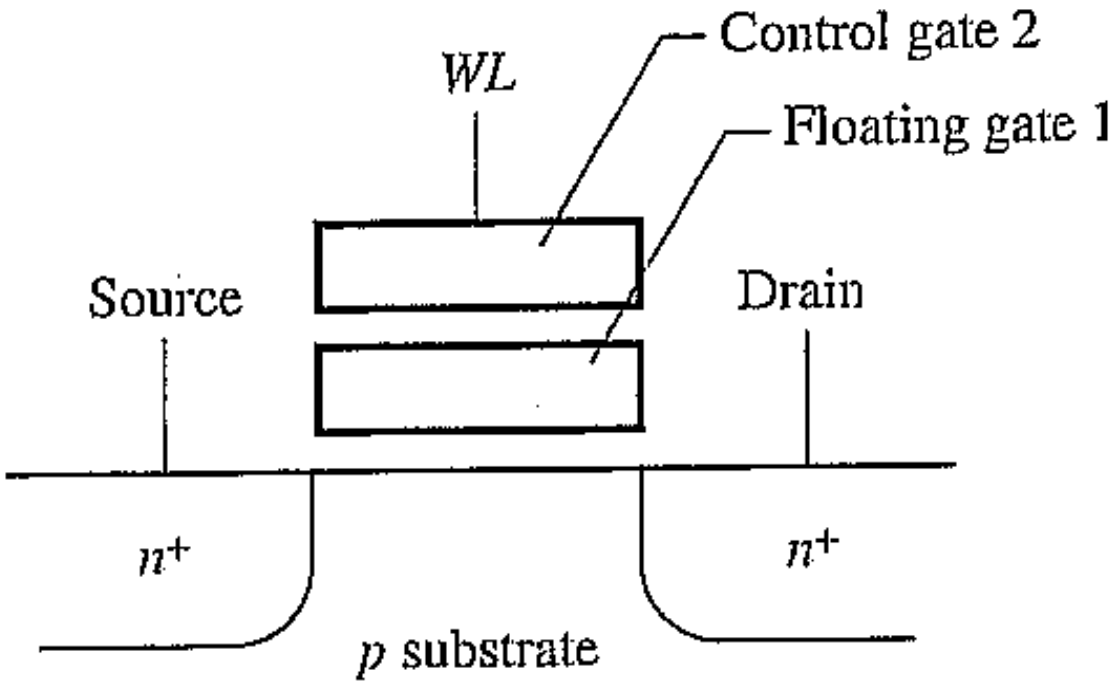
Read / Write Non-volatile Memory
- USB memory sticks
- Digital cameras
- MP3 players
- …

10^6 R/W operations

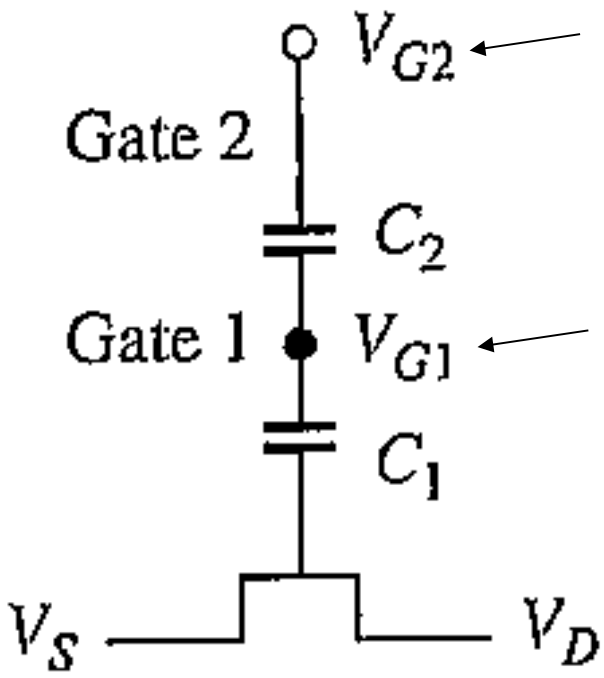
1T cell implementation of E^2PROM

Two architectures
- NAND
- NOR

Dual-Gate Structure



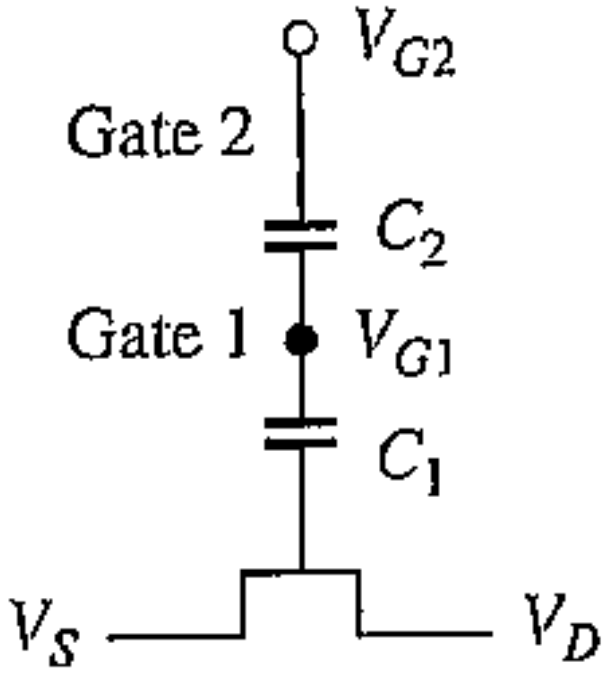
Schematic Representation



Applied Gate Voltage

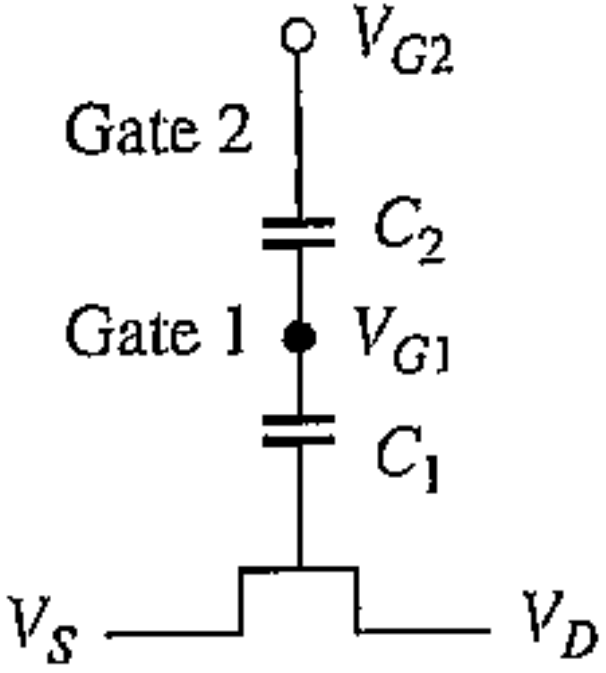
Gate Voltage Which Controls Device Characteristics

Capacitive Divider at Gate



$$\Delta V_{G1} = \Delta V_{G2} \quad \frac{C_2}{C_1 + C_2}$$

To Turn The Device On

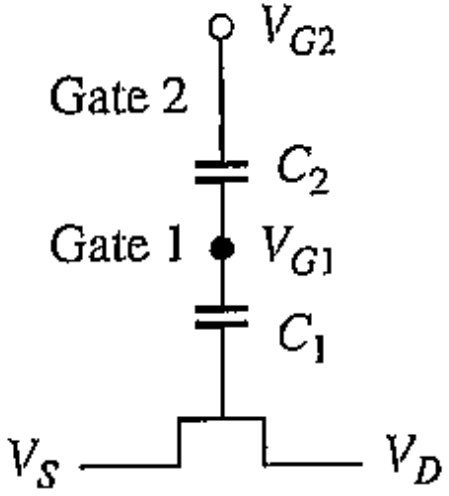


$$V_{G1} > V_t$$

$$\Rightarrow V_{G1} = V_{G2} \frac{C_2}{C_1 + C_2} > V_t$$

$$\Rightarrow V_{G2} > V_t \frac{C_1 + C_2}{C_2}$$

Device Operation



$$V_{G1} > V_t$$

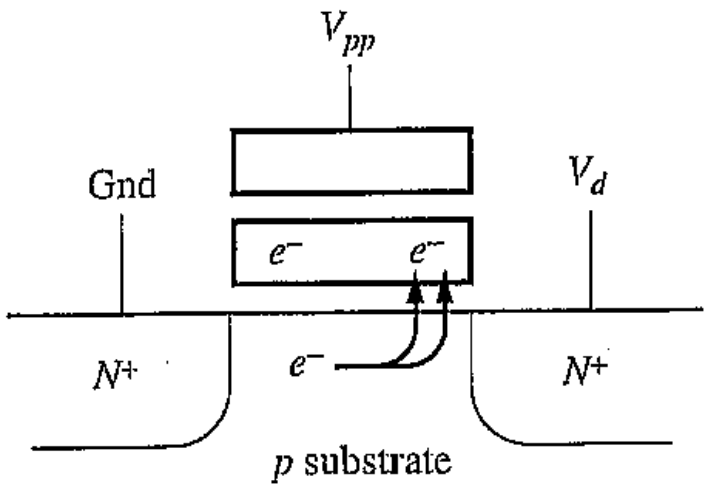
$$\Rightarrow V_{G1} = V_{G2} \frac{C_2}{C_1 + C_2} > V_t$$

$$\Rightarrow V_{G2} > V_t \frac{C_1 + C_2}{C_2}$$

Device operation relies on the ability to store and remove charge from the floating gate → namely node V_{G1}

Write Process

Write process relies on “Hot Carrier Injection” (HCL)

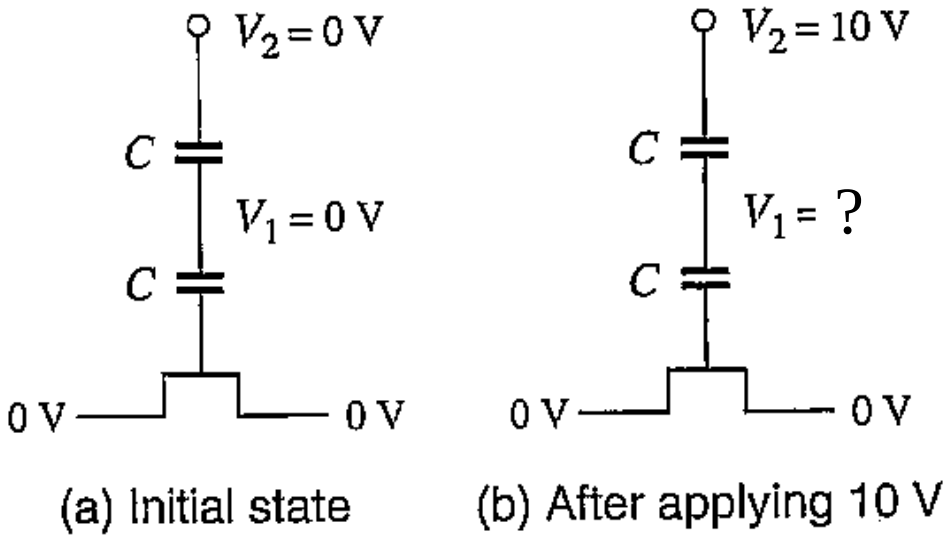


- Raise V_{G2} and V_{DS} >> VDD
- Avalanche breakdown of Drain/Substrate Junction
- Electrons are injected through the gate oxide onto the floating gate
- Self-terminating process because V_{G1} decreases as electrons are injected onto the floating gate

Example

Assume C1 = C2 = C

VDD = 5V, V_T = 1V

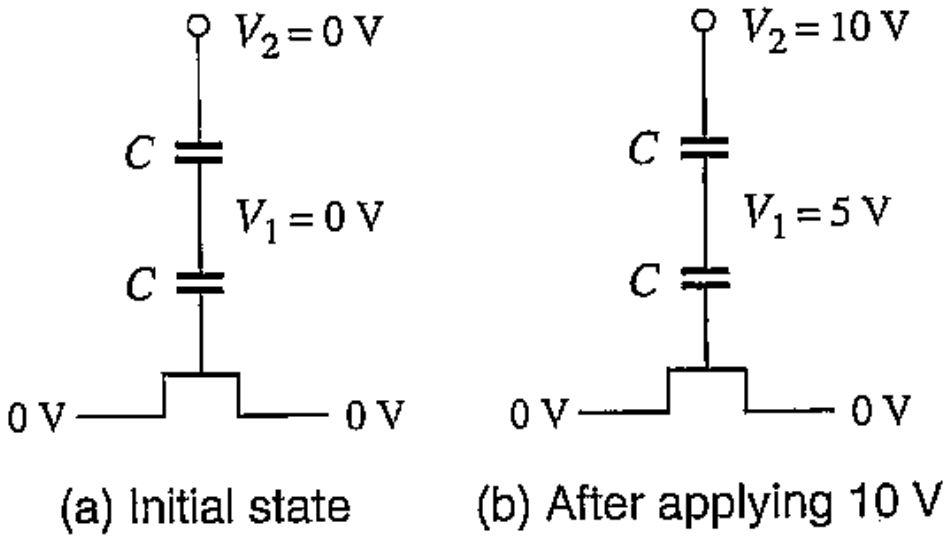


(a) Initial state (b) After applying 10 V

Example

Assume C1 = C2 = C

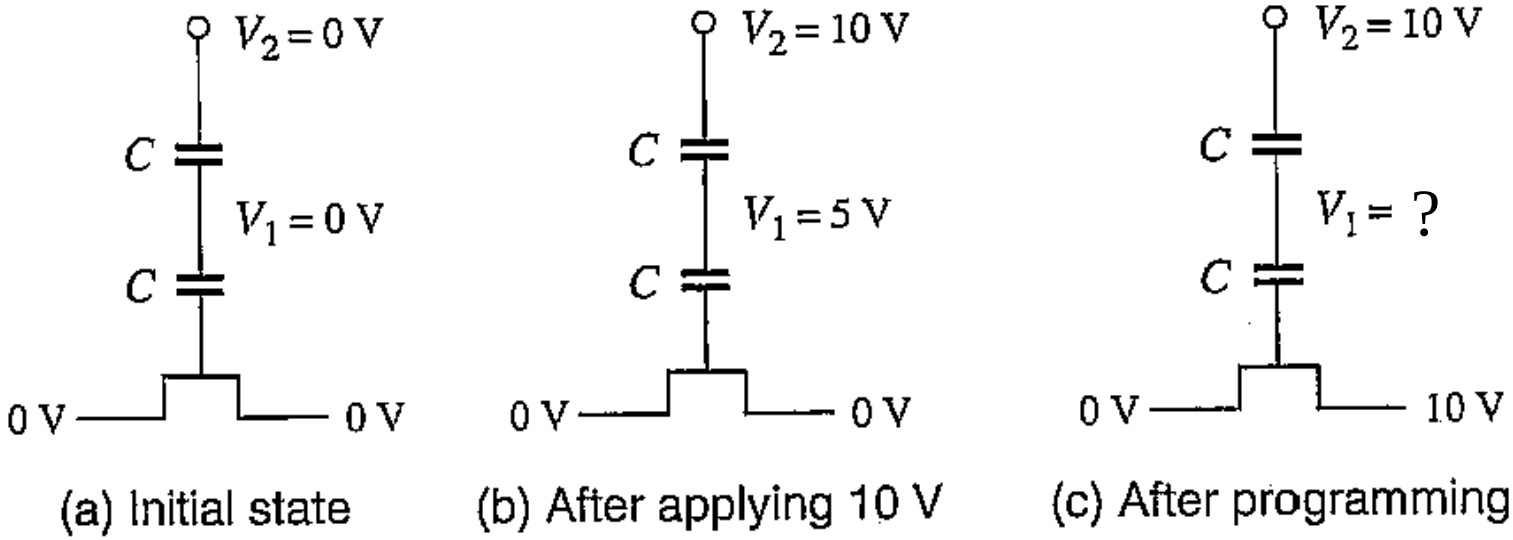
VDD = 5V, V_T = 1V



(a) Initial state (b) After applying 10 V

Example

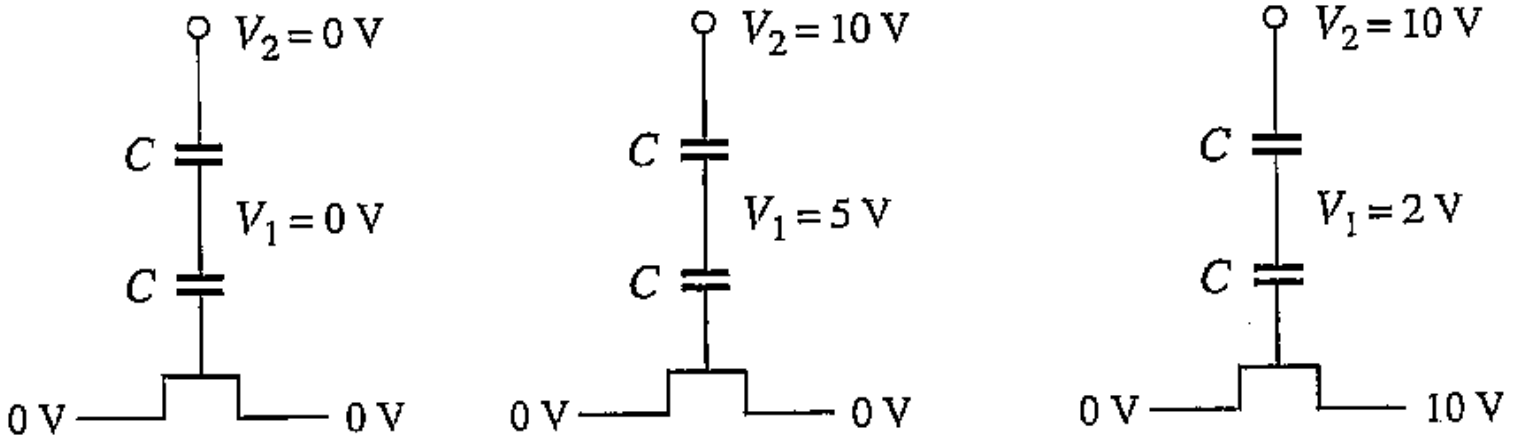
Assume C1 = C2 = C
VDD = 5V, V_T = 1V



Program by applying V_{DS} = 10V

Example

Assume C1 = C2 = C
VDD = 5V, V_T = 1V

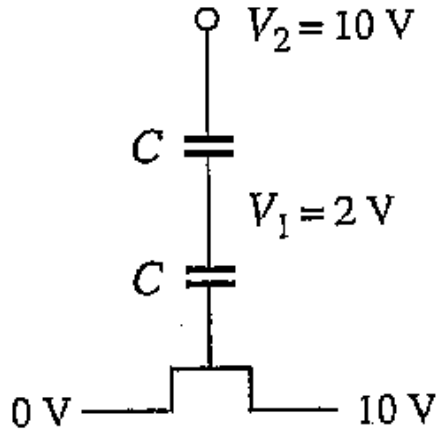


Program by applying V_{DS} = 10V
Assume V_{G1} drops to 2V

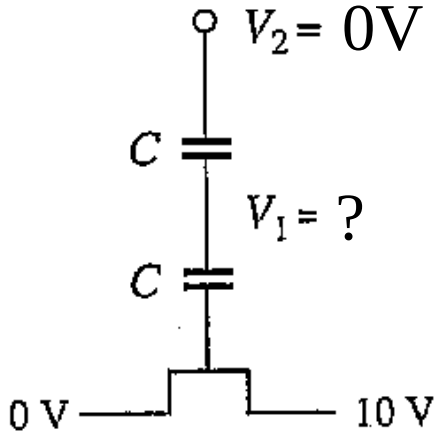
Example

Assume C1 = C2 = C

VDD = 5V, V_T = 1V



(c) After programming

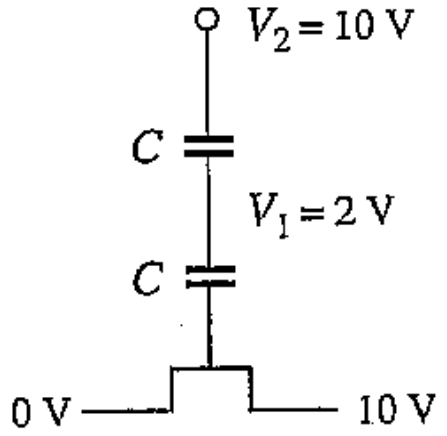


Reduce gate voltage V_{G2} = 0V

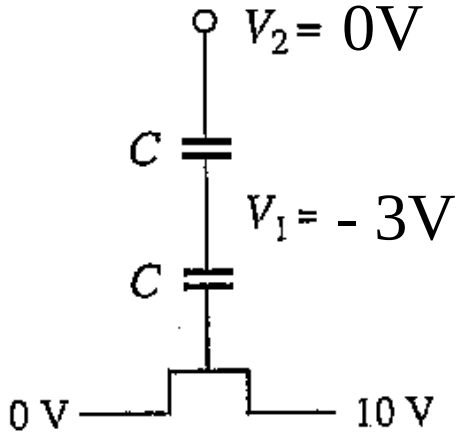
Example

Assume C1 = C2 = C

VDD = 5V, V_T = 1V



(c) After programming

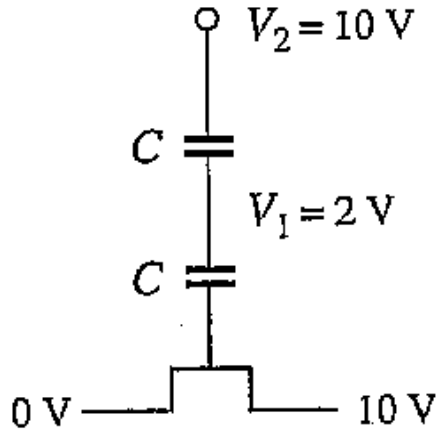


Reduce gate voltage V_{G2} = 0V

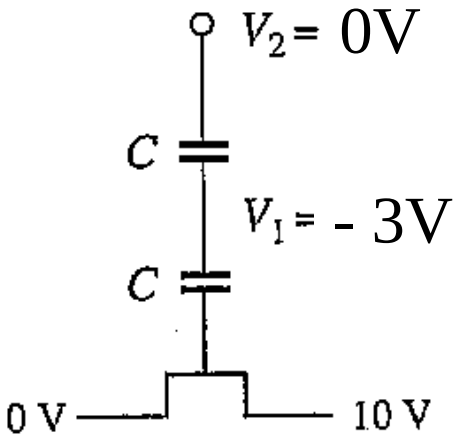
Example

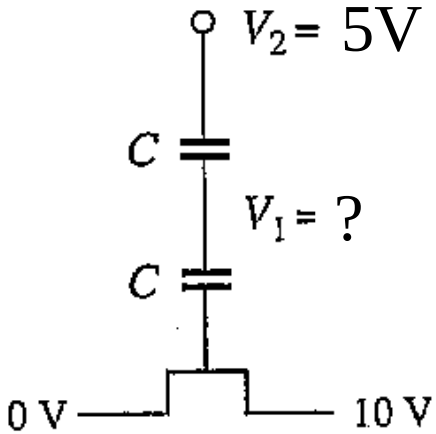
Assume C1 = C2 = C

VDD = 5V, V_T = 1V



(c) After programming

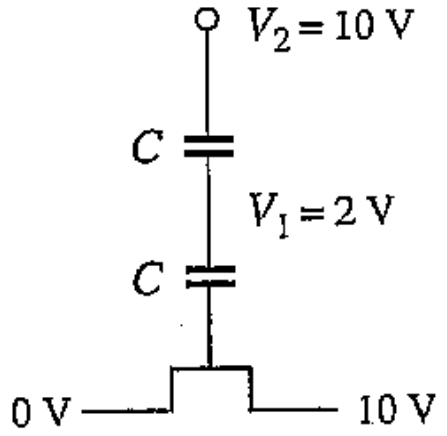




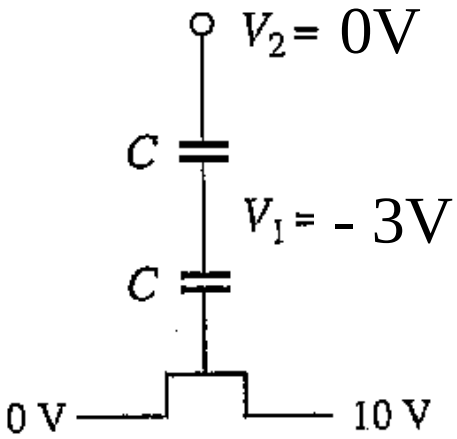
Raise gate voltage V_{G2} = VDD

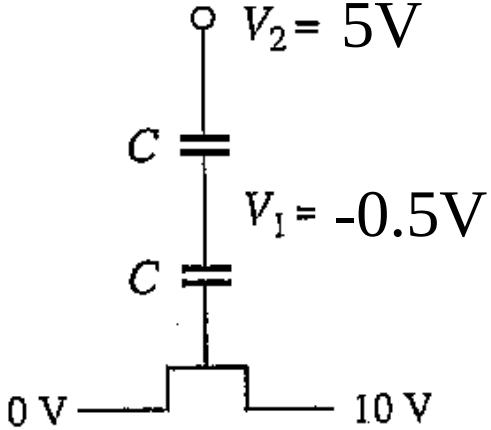
Example

Assume C1 = C2 = C
VDD = 5V, V_T = 1V



(c) After programming

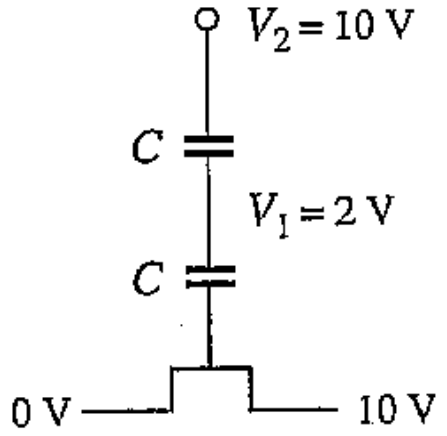




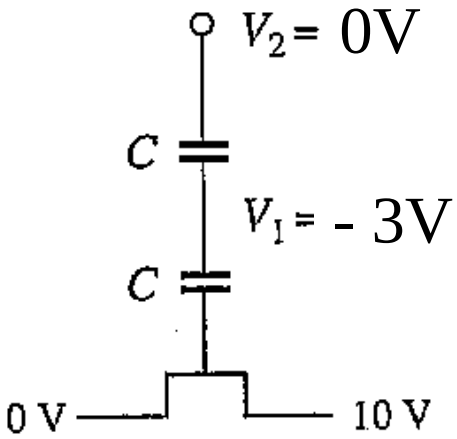
Raise gate voltage V_{G2} = VDD
TRANSISTOR IS STILL OFF!!
Have written a “1”

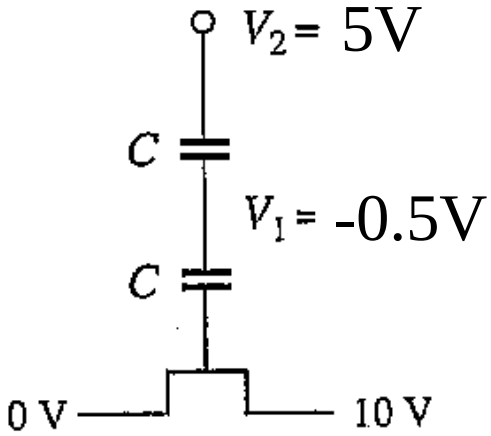
Example

Assume C1 = C2 = C
VDD = 5V, V_T = 1V



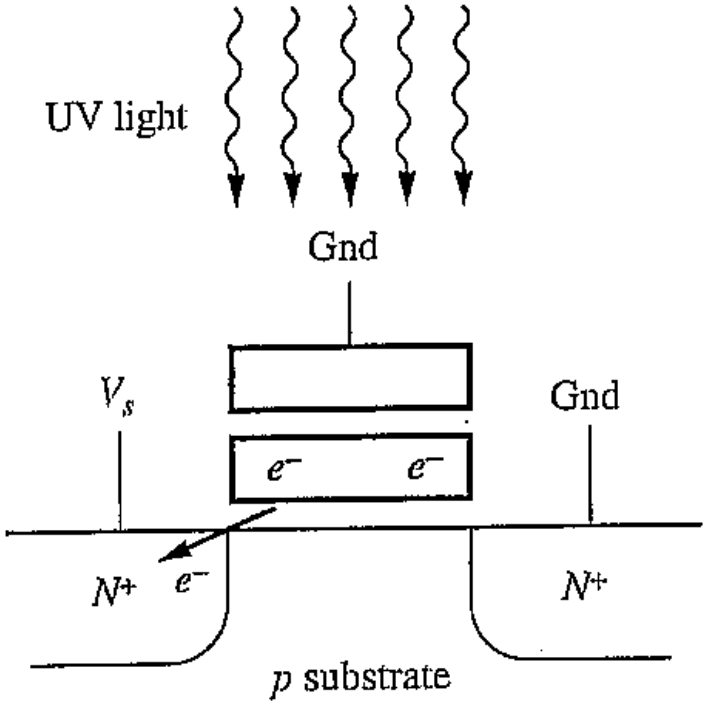
(c) After programming





Raise gate voltage V_{G2} = VDD
TRANSISTOR IS STILL OFF!!
To turn the transistor on: V_{G2} = 8V

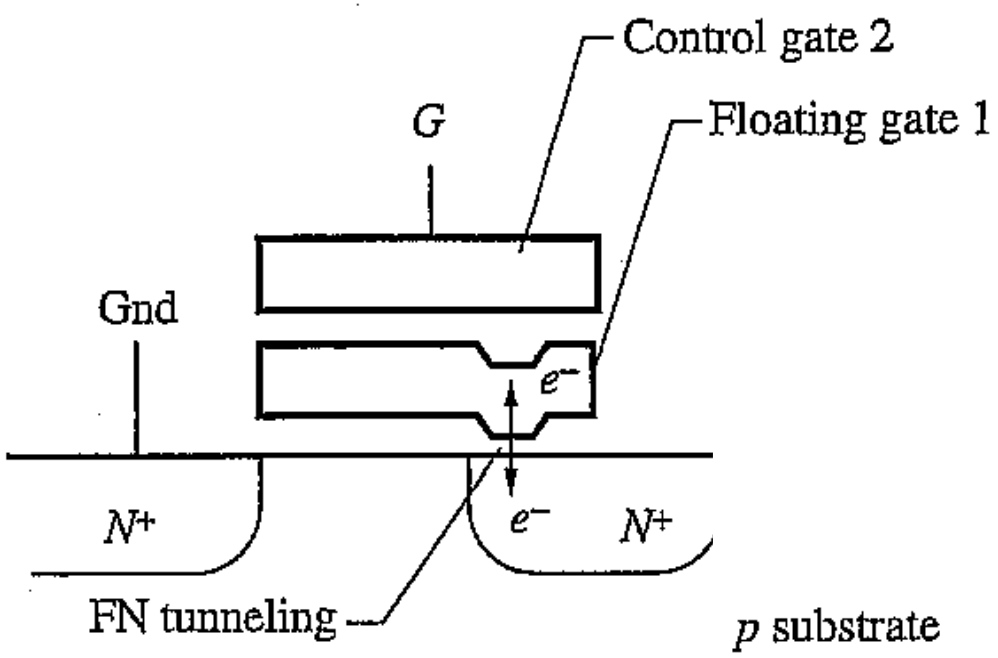
Different Methods to Erase



(b) Erase process —UV light

In a conventional EPROM, UV light is used generate e/h pairs in the SiO_2 dielectric making the dielectric conductive

Different Methods to Erase

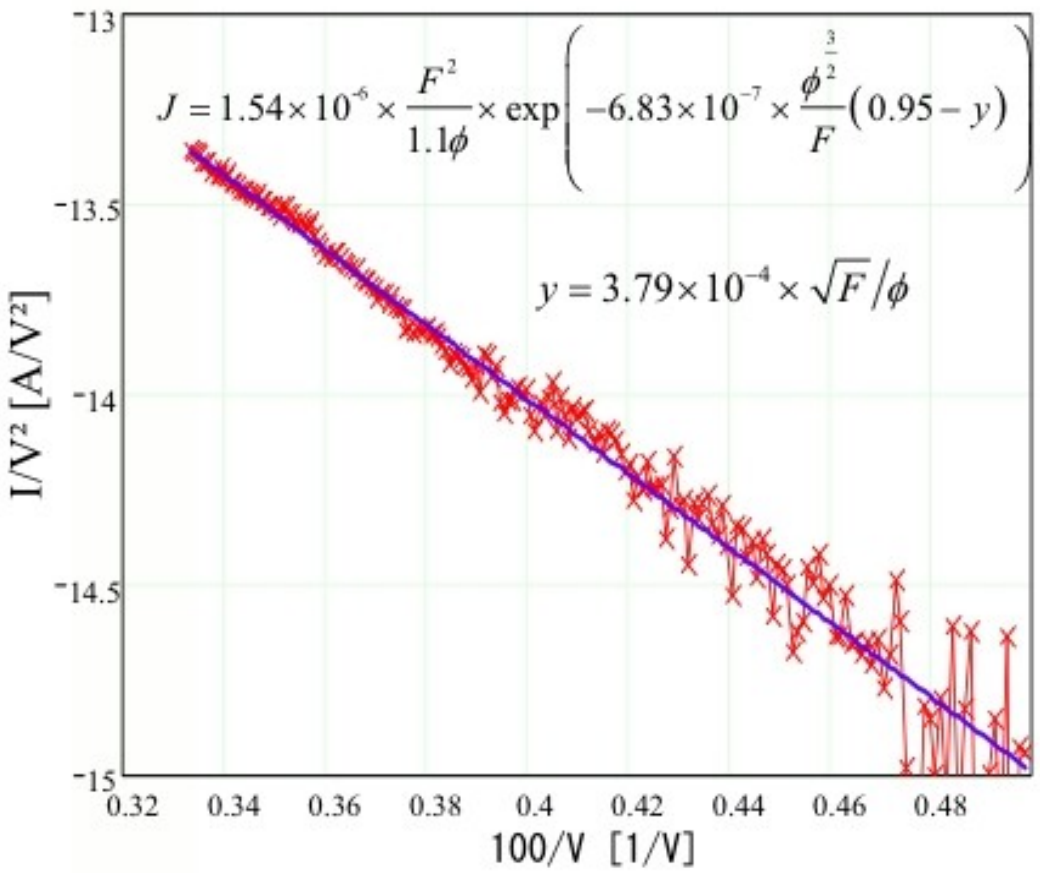


In a E^2PROM, the charge is removed by "Fowler-Nordheim" (FN) transport.

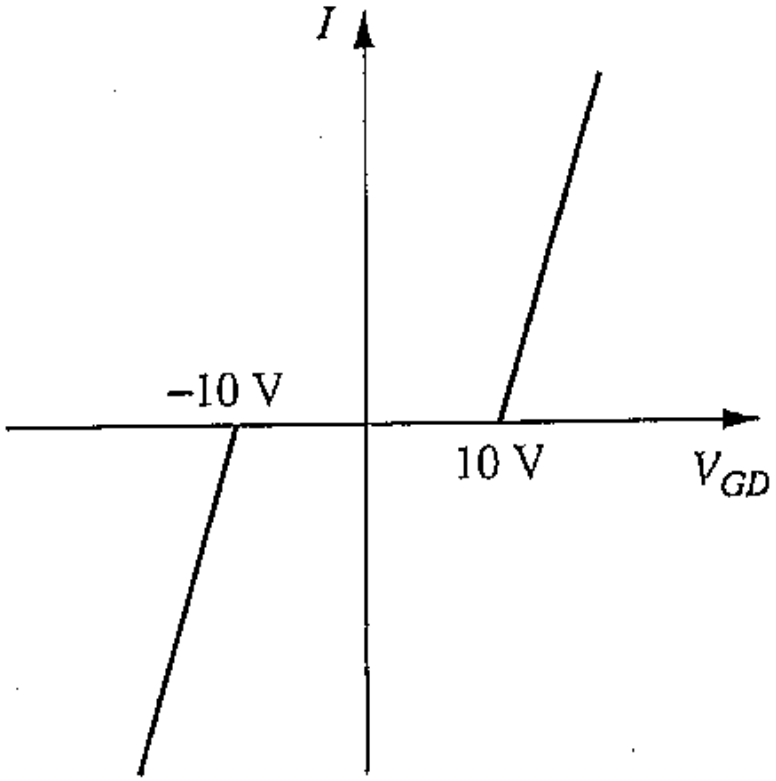
When the field across the oxide exceed ~10 MV/cm, electrons can tunnel across the oxide

Fowler-Nordheim Tunneling

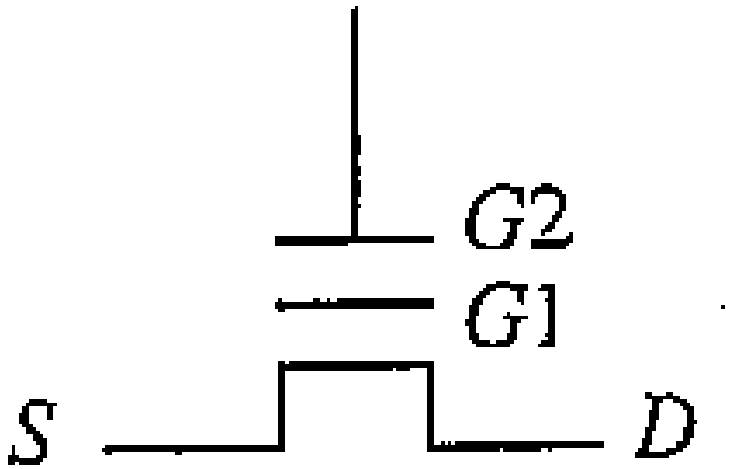
$$j_{FN} = BE^2 e^{\left(-\frac{b}{E}\right)}$$



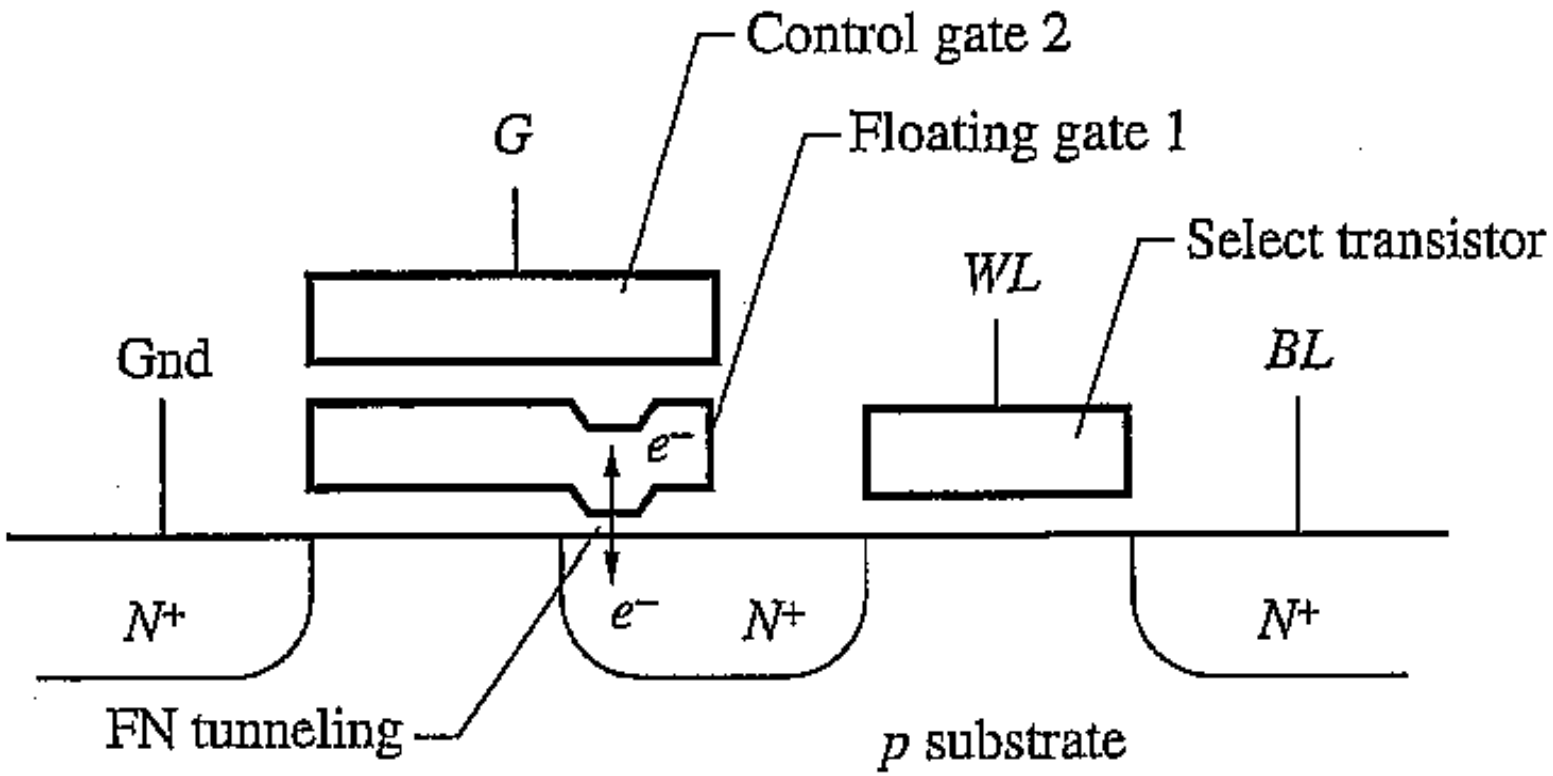
FN Tunneling – 10nm SiO_2



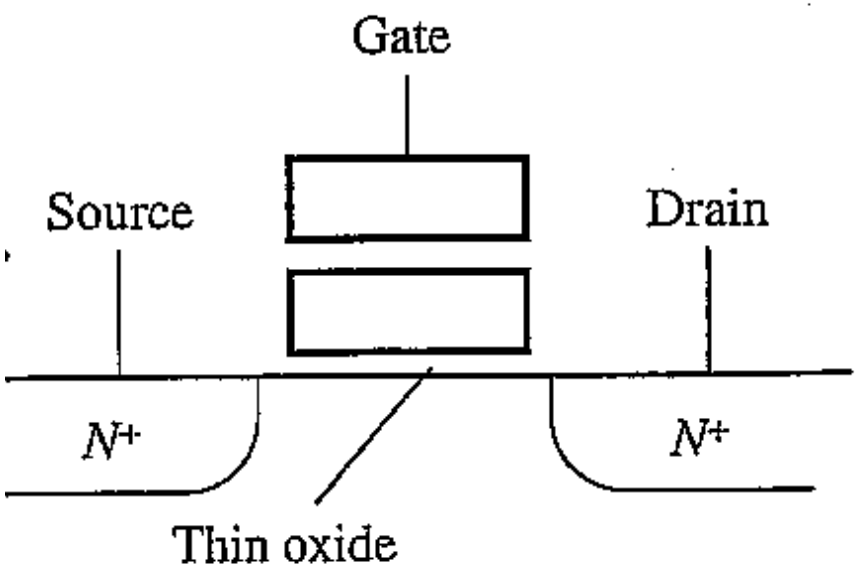
Symbol



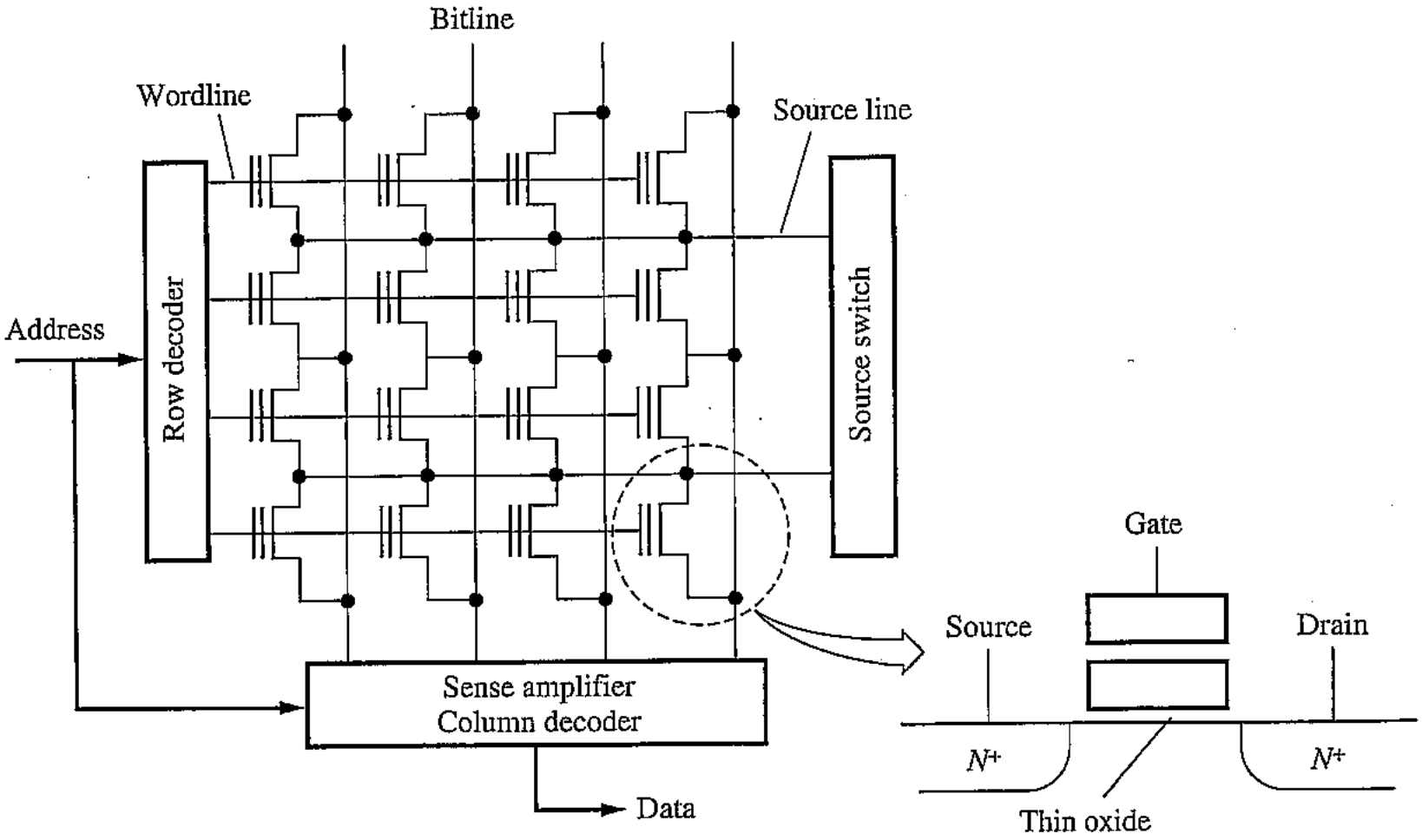
2T E²PROM Cell



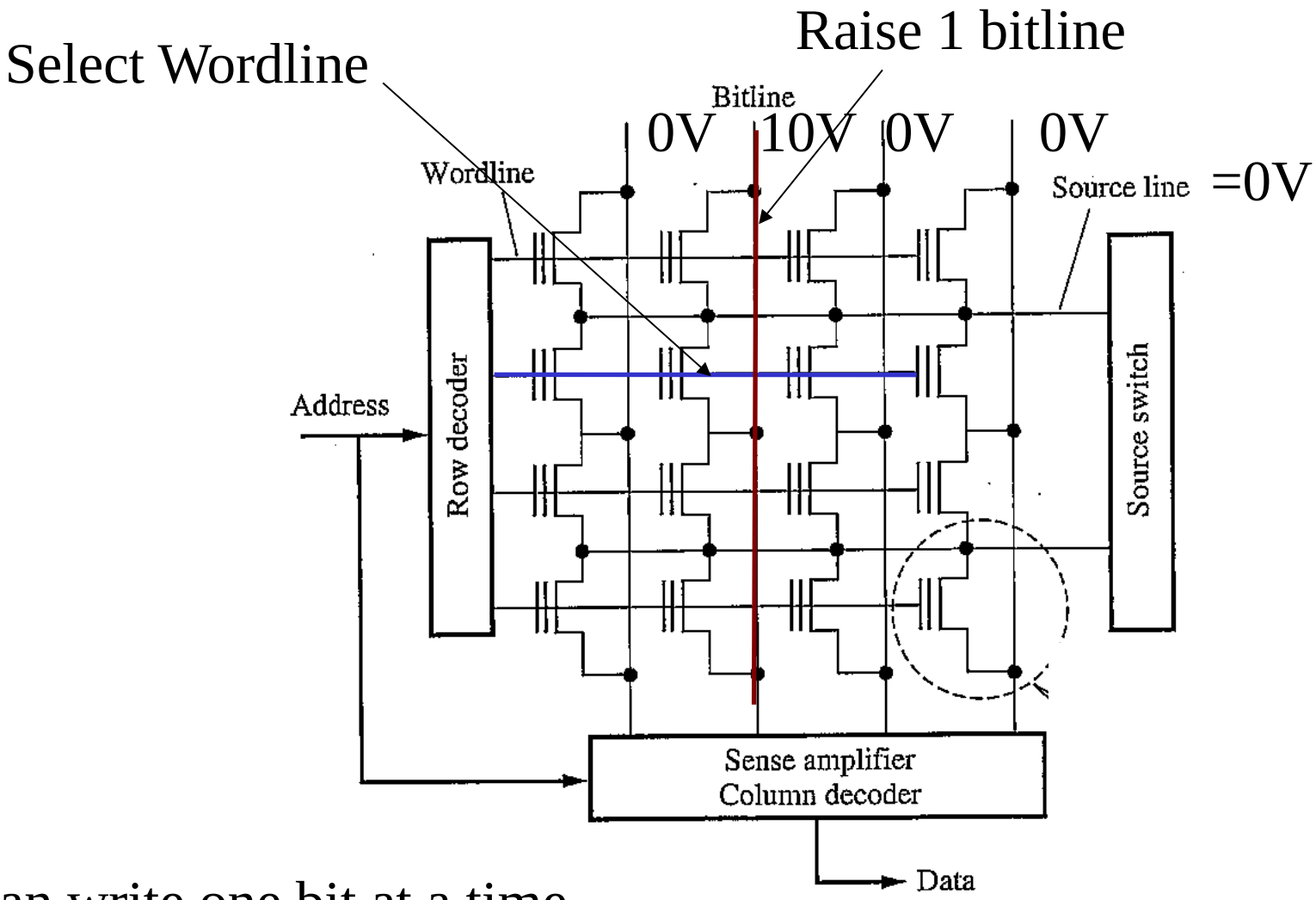
1T Cell - FLASH



NOR Array


Bitline
Wordline
Source line
Address
Row decoder
Source switch
Sense amplifier
Column decoder
Data
Gate
Source
Drain
N+
N+
Thin oxide

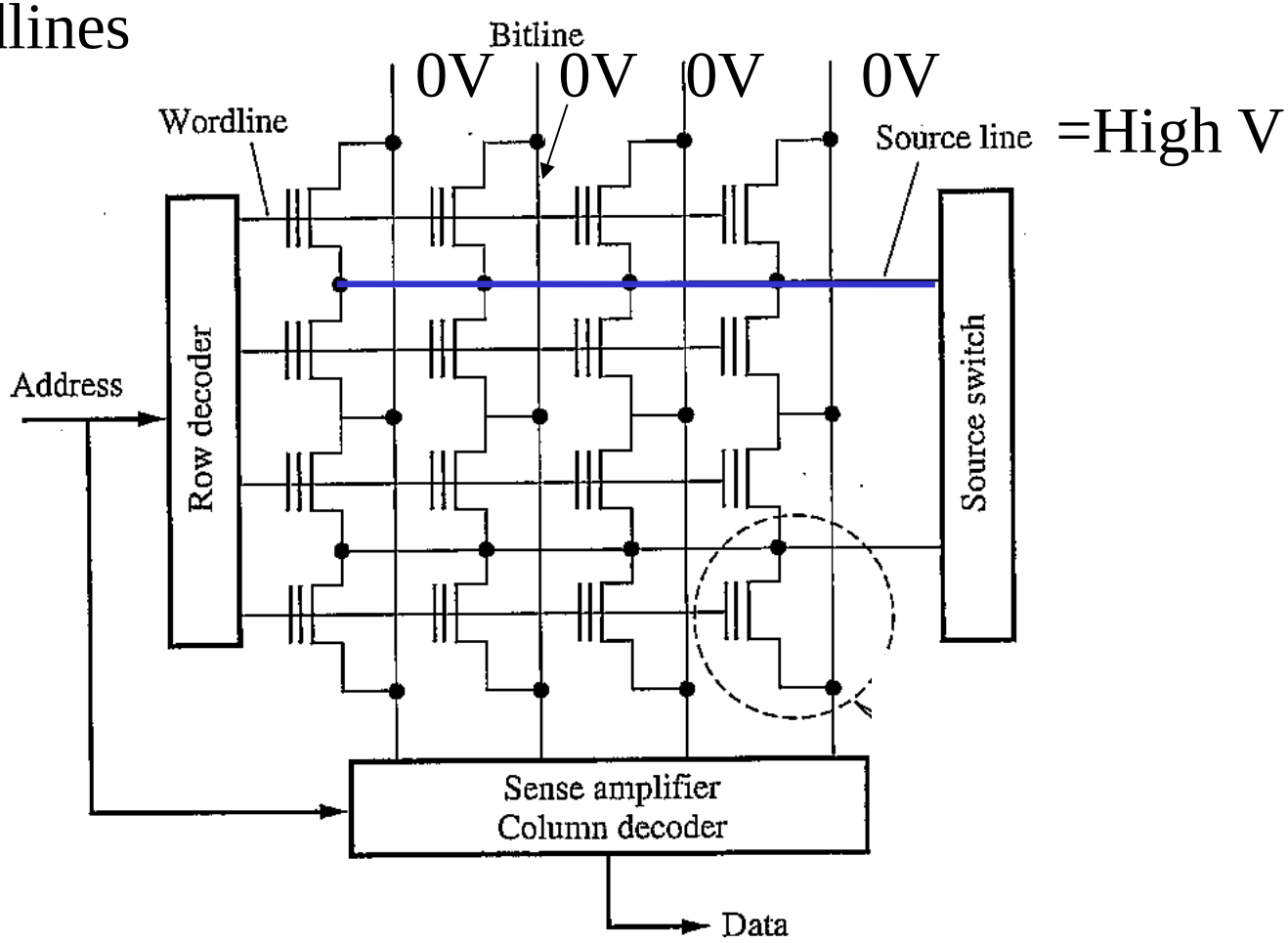
Write – “hot carrier injection”



Can write one bit at a time

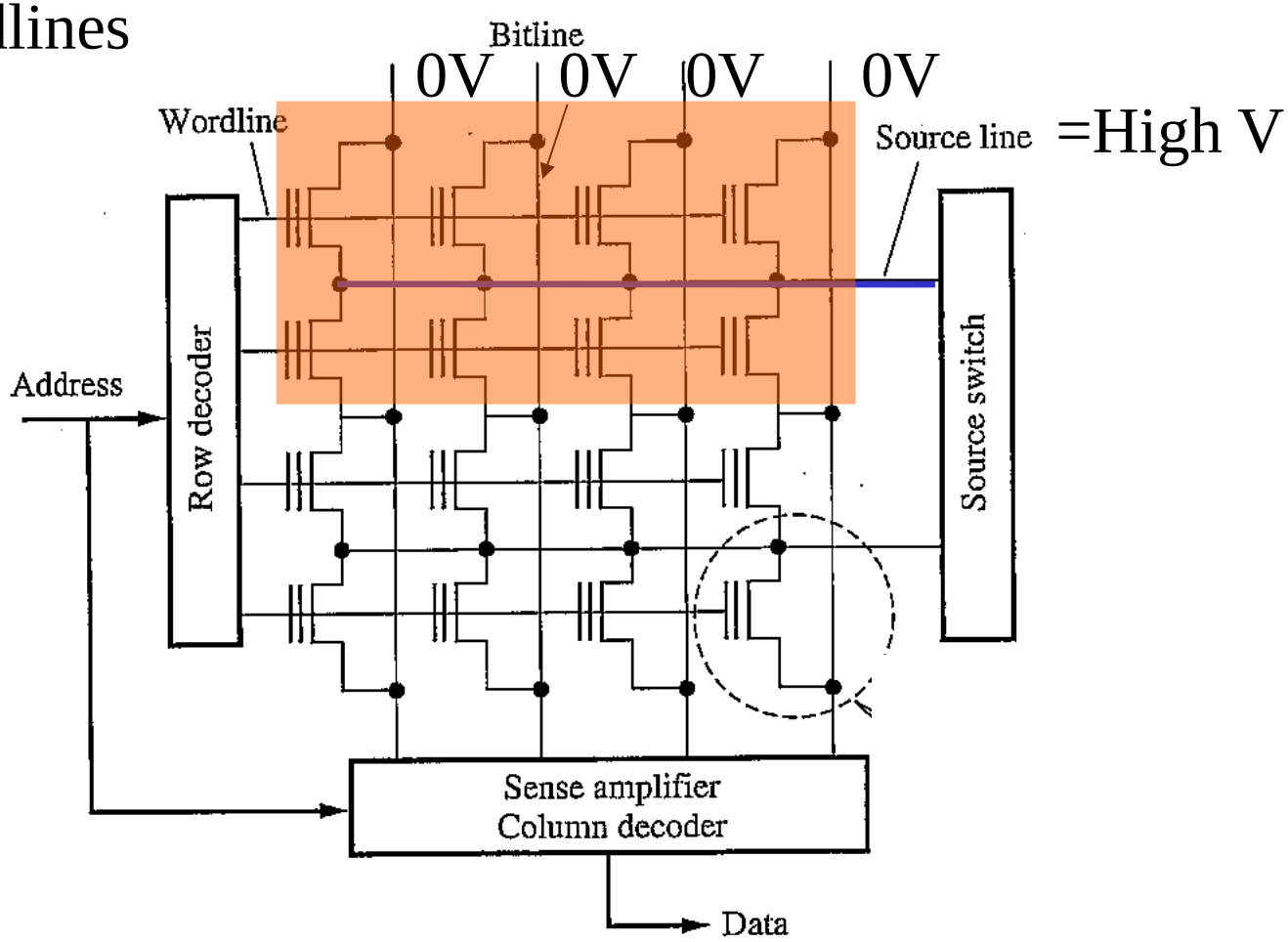
Erase – “Fowler-Nordheim Tunneling”

0V on Wordlines



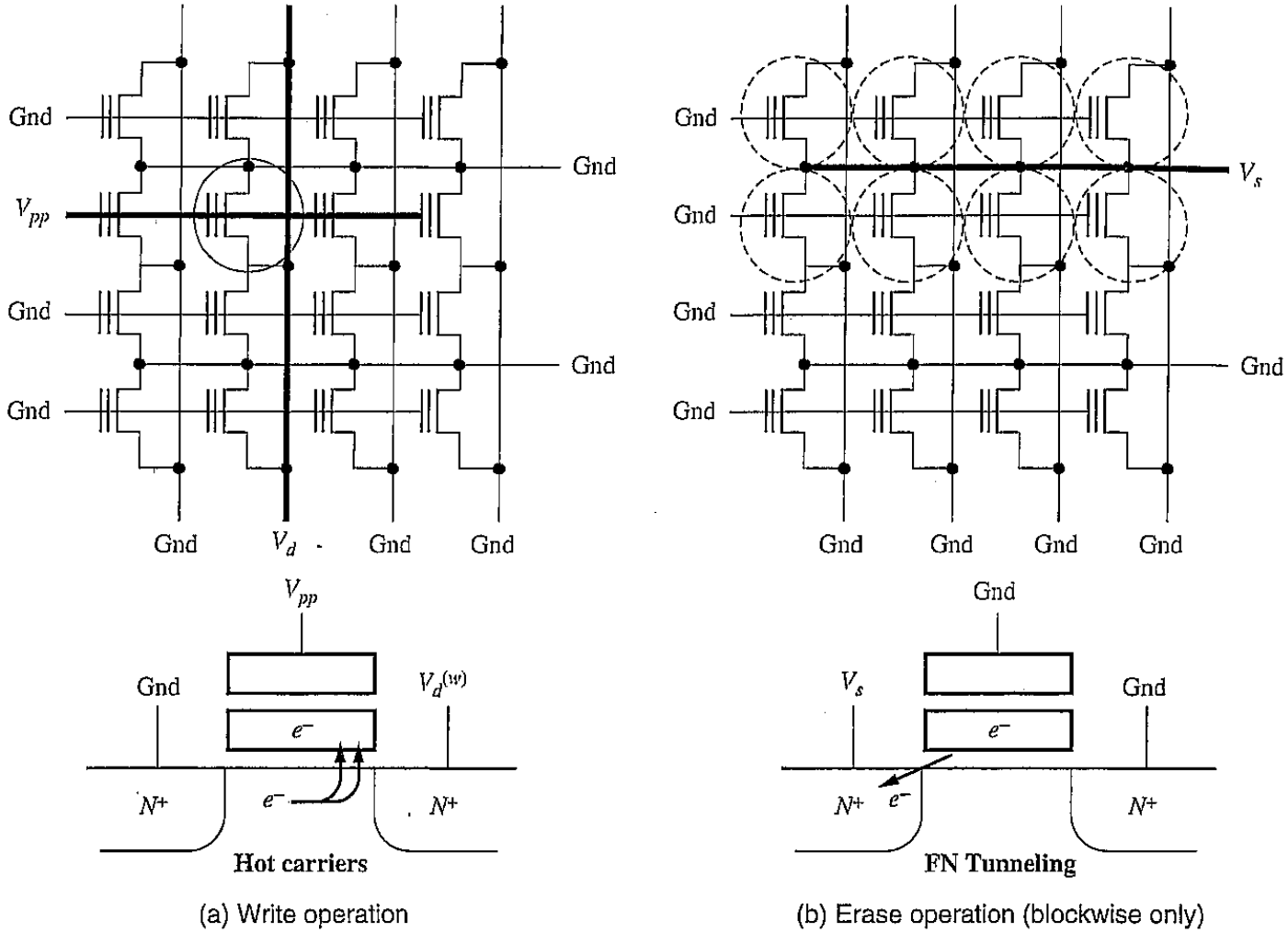
Erase – “Fowler-Nordheim Tunneling”

0V on Wordlines

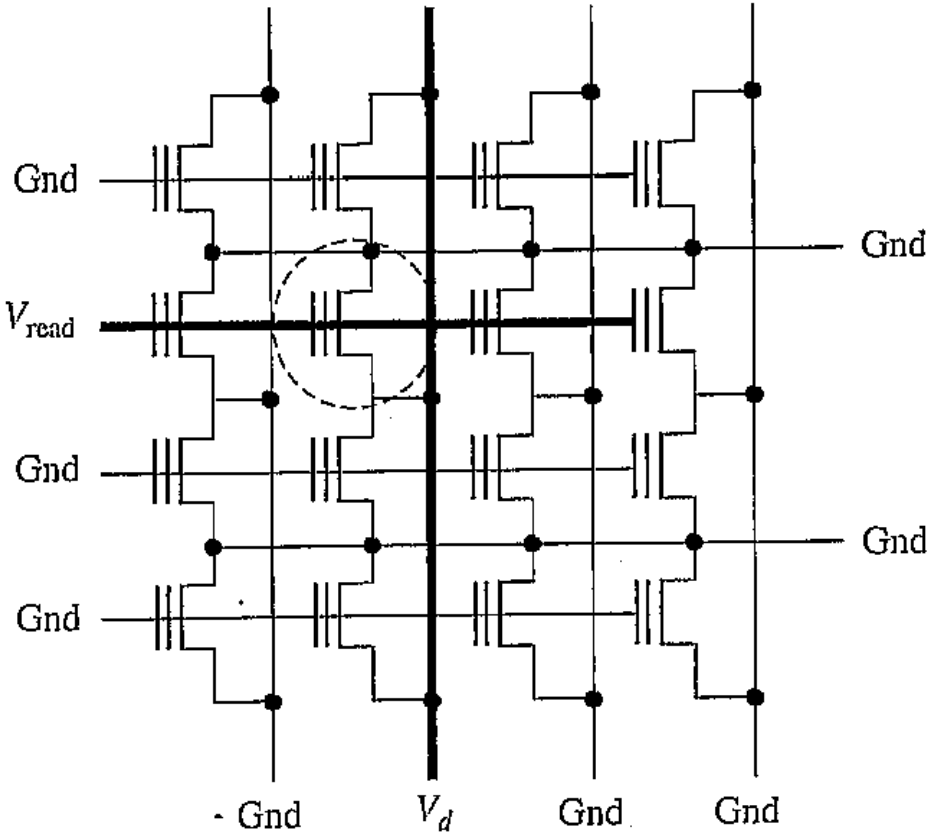


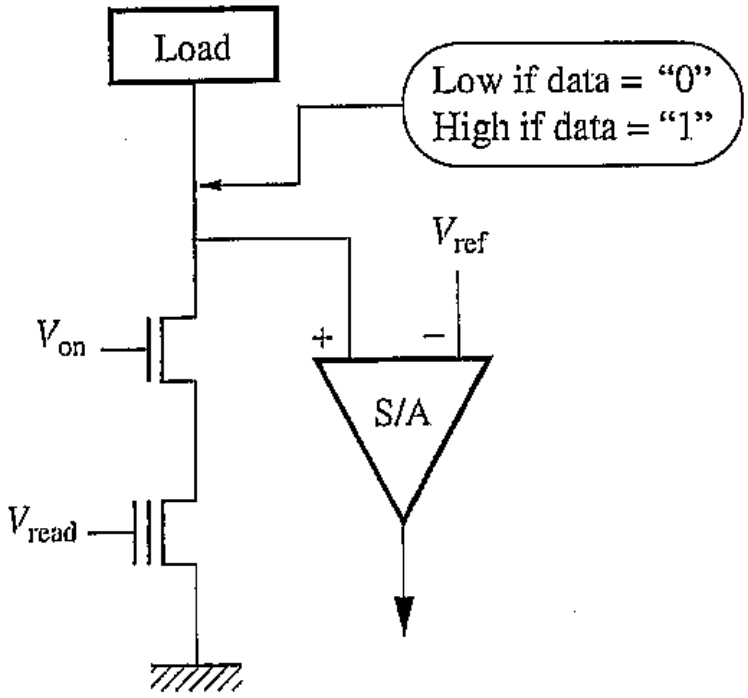
All transistors connected to the source line are erased simultaneously → FLASH

Write / Erase


Gnd
Gnd
V_{pp}
Gnd
Gnd
Gnd
Gnd
V_d
Gnd
Gnd
V_{pp}
Gnd
V_d(w)
e^-
e^-
N+
N+
Hot carriers
(a) Write operation
Gnd
Gnd
V_s
Gnd
Gnd
Gnd
Gnd
Gnd
Gnd
Gnd
Gnd
V_s
Gnd
e^-
e^-
N+
N+
FN Tunneling
(b) Erase operation (blockwise only)

Read Operation



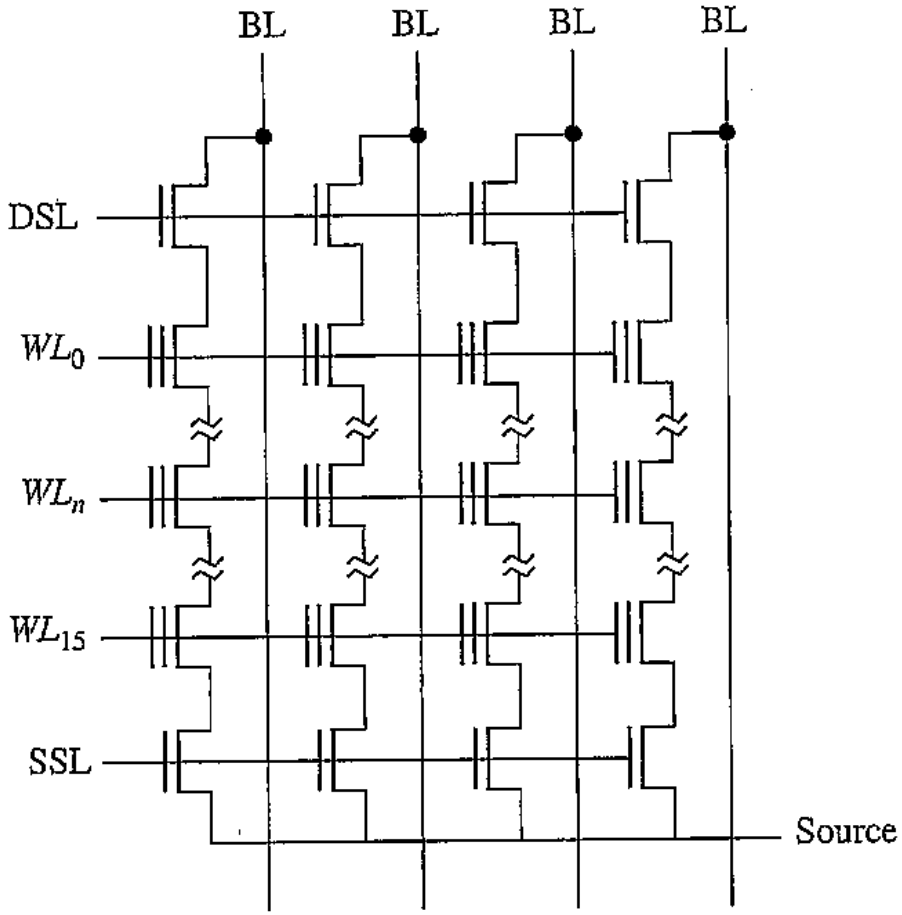


Precharge bitline

Source line → Gnd

Enable wordline

NAND Flash Array



NAND vs NOR Arrays

	Write	Read	Erase	Density
NAND	Fast	Slow	Fast	High
NOR	Slow	Fast	Slow	Low

Issues

Read / Write Endurance

- > 10^6 Operations

Redundancy for improved yield

Multi-Bit Cell Architectures – in common practice today!