Lecture 12: Memory (SRAM, DRAM, Flash)

Jangwoo Kim (Seoul National University)
jangwoo@snu.ac.kr

Read-Only Memories

- Mask ROM
  - Programmed permanently during manufacturing process
- Programmable ROM
  - PROM
    - programmed by fusing or anti-fusing process
    - the process is irreversible
  - Erasable PROM
    - UV EPROM
    - EEPROM

Random Access Memories

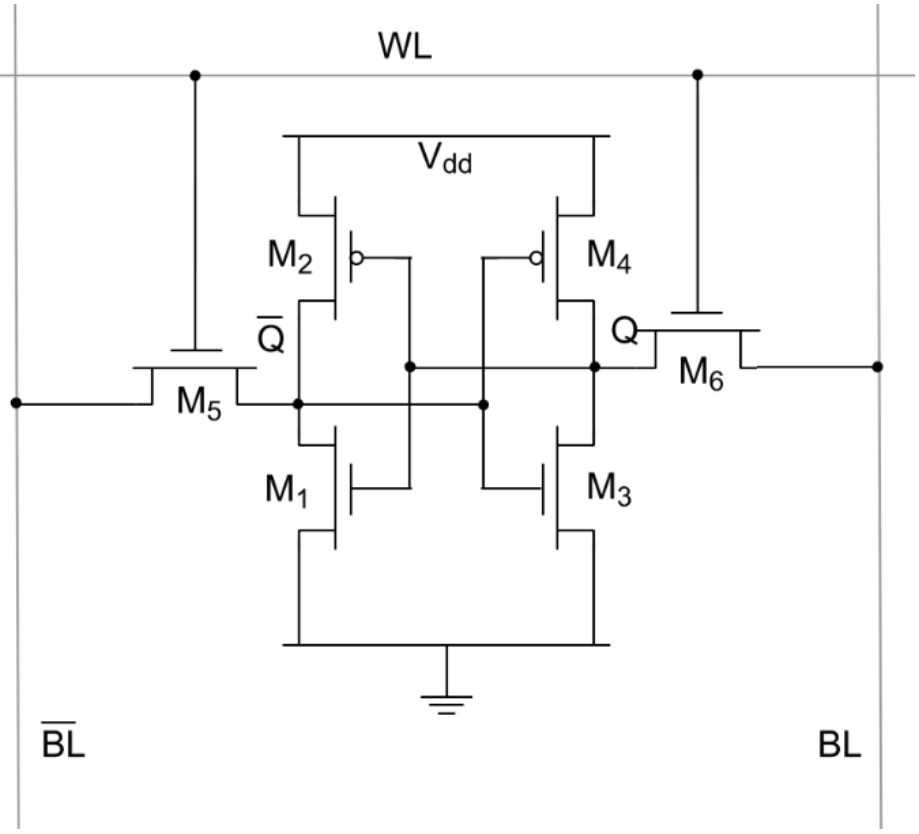
- Transistor efficient methods for implementing storage elements
- Small RAM: 256 words by 4-bit
- Large RAM: 1 billion words by 1-bit
- Static RAMs and Dynamic RAMs

SRAM

(Static Random Access Memory)

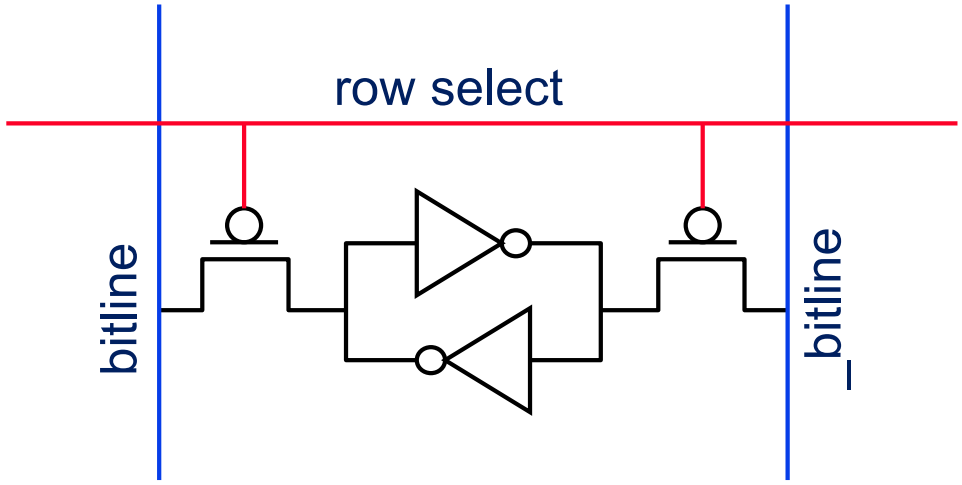
Static RAM (SRAM) (1/2)

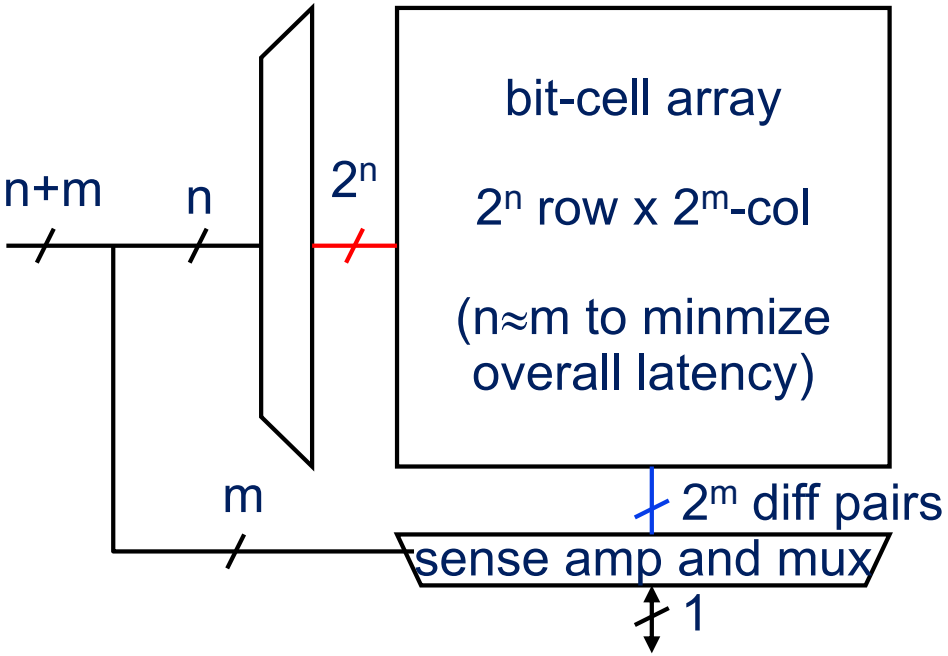
- Static Random Access Memory
  - Cache
  - Register file
  - …
- Read
  - Sense bitline difference
- Write
  - Overwrite bitlines using strong signals



1 Cell = 6 transistors (typical) 4 (two invertors) + 2 (switches)

Static RAM (SRAM) (2/2)





Read Sequence

(assume bitlines are precharged)

1. Address decode
2. Drive row select
3. Selected bit-cells drive bitlines
   (entire row is read together)
4. Diff. sensing and column select
   (data is ready)
5. Precharge all bitlines
   (for next read or write)

How do you write select columns?

Access latency dominated by steps 2, 3

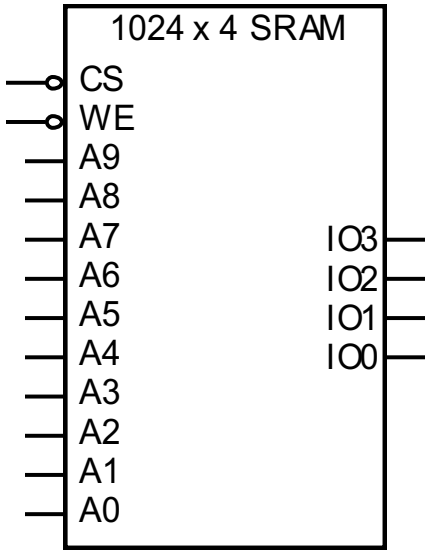
Cycle time dominated by steps 2, 3, 5

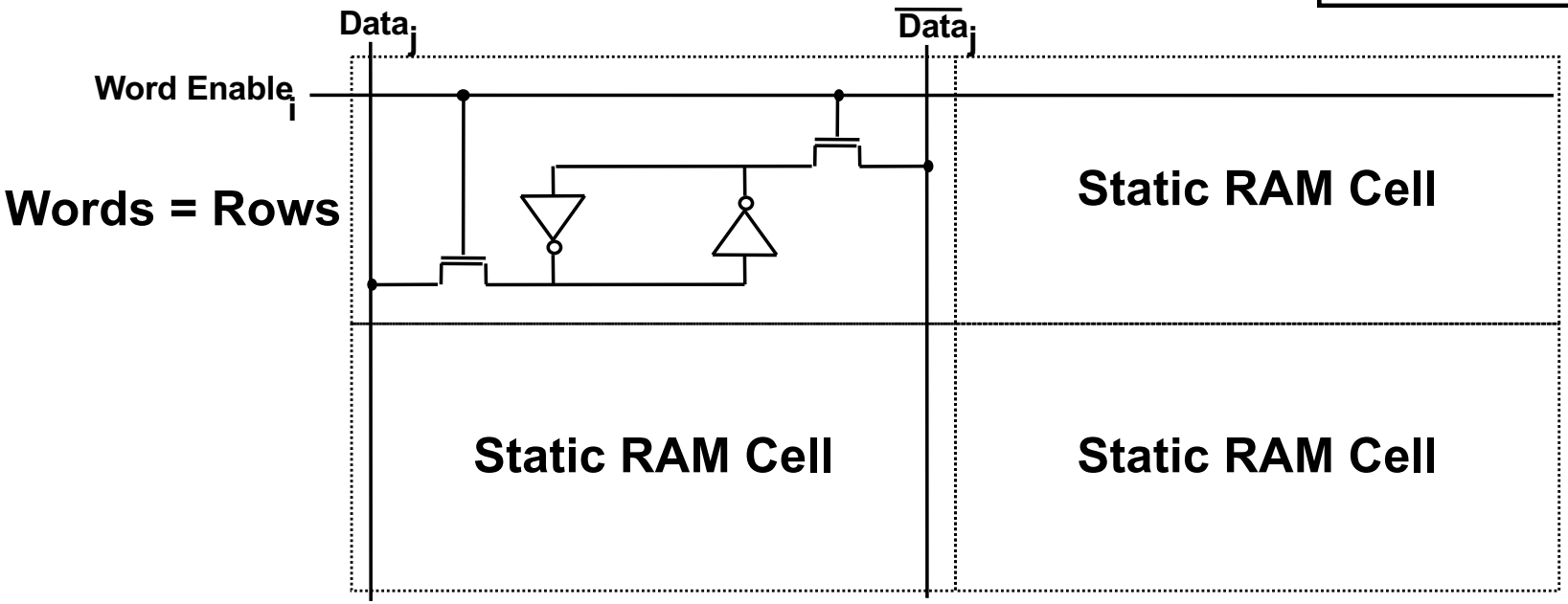
- Step 2 proportional to 2^m
- Step 3 and 5 proportional to 2^n

usually encapsulated by synchronous (sometime pipelined) interface logic

Static RAMs

- We will discuss a 1024 x 4 organization
  - **Chip Select Line (active lo)**
  - **Write Enable Line (active lo)**
  - **10 Address Lines**
  - **4 Bidirectional Data Lines**



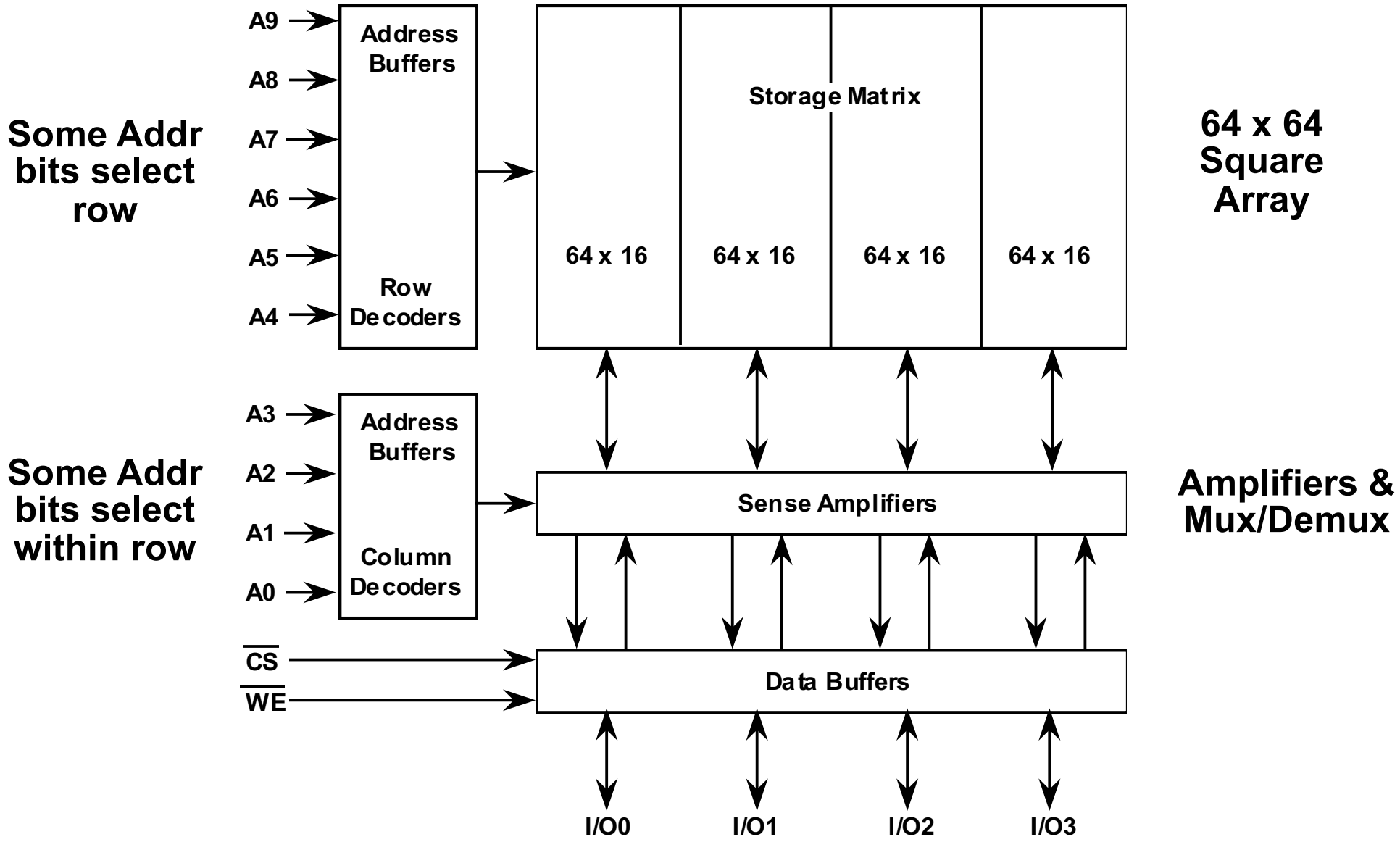


Columns = Bits (Double Rail Encoded)

Static RAMs

RAM Organization

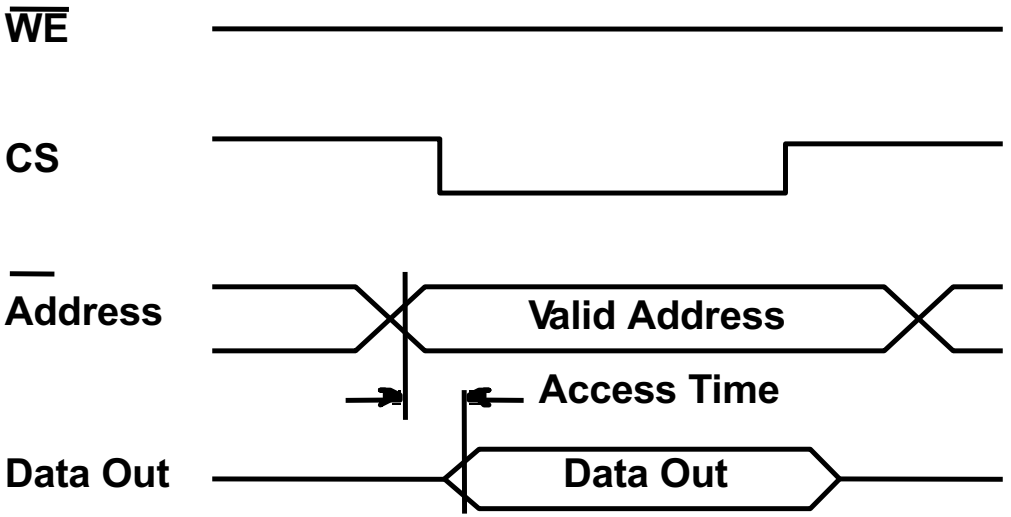
Long thin layouts are not the best organization for a RAM



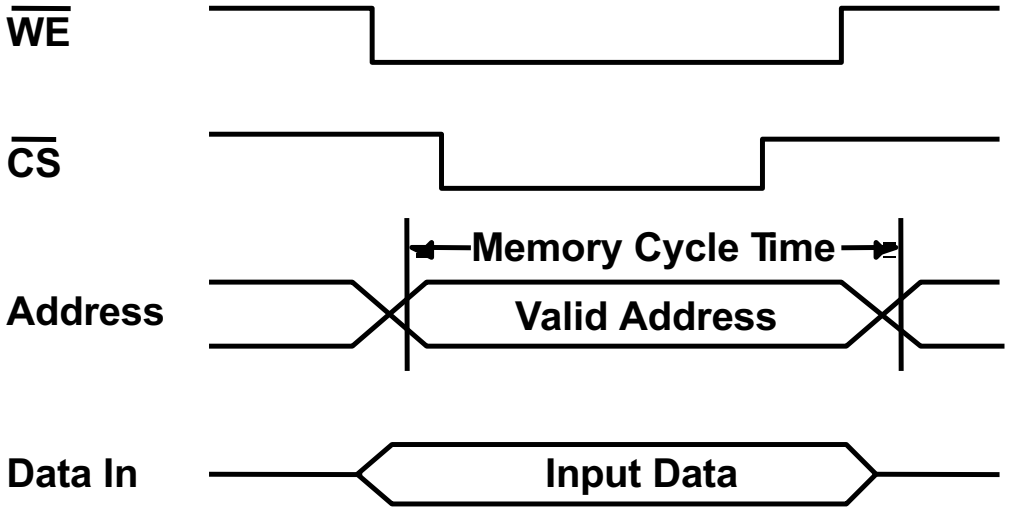
Static RAMs

RAM Timing

Simplified Read Timing

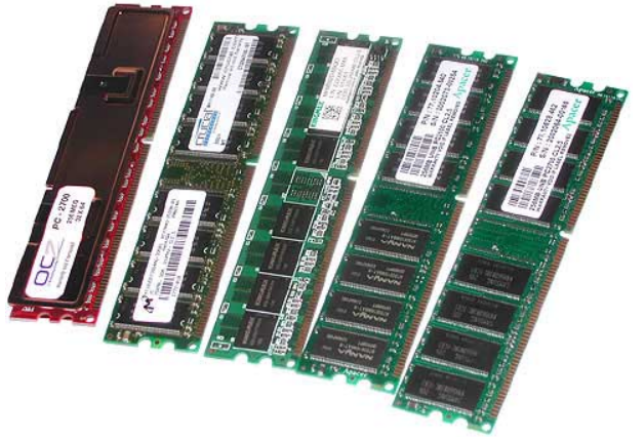


Simplified Write Timing

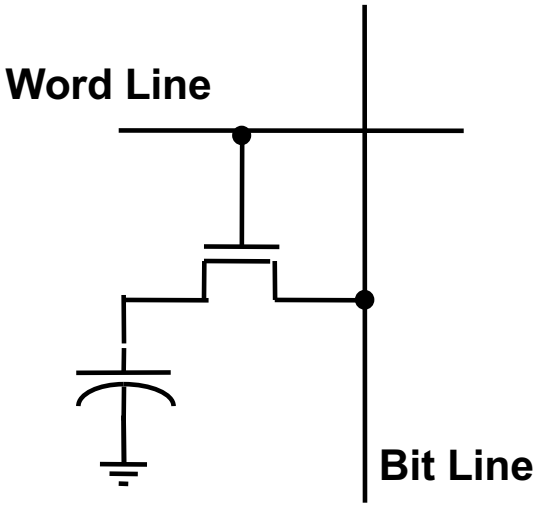


DRAM

(Dynamic Random Access Memory)



Dynamic RAMs



1 Transistor (+ capacitor) memory element

Read: Assert Word Line, Sense Bit Line

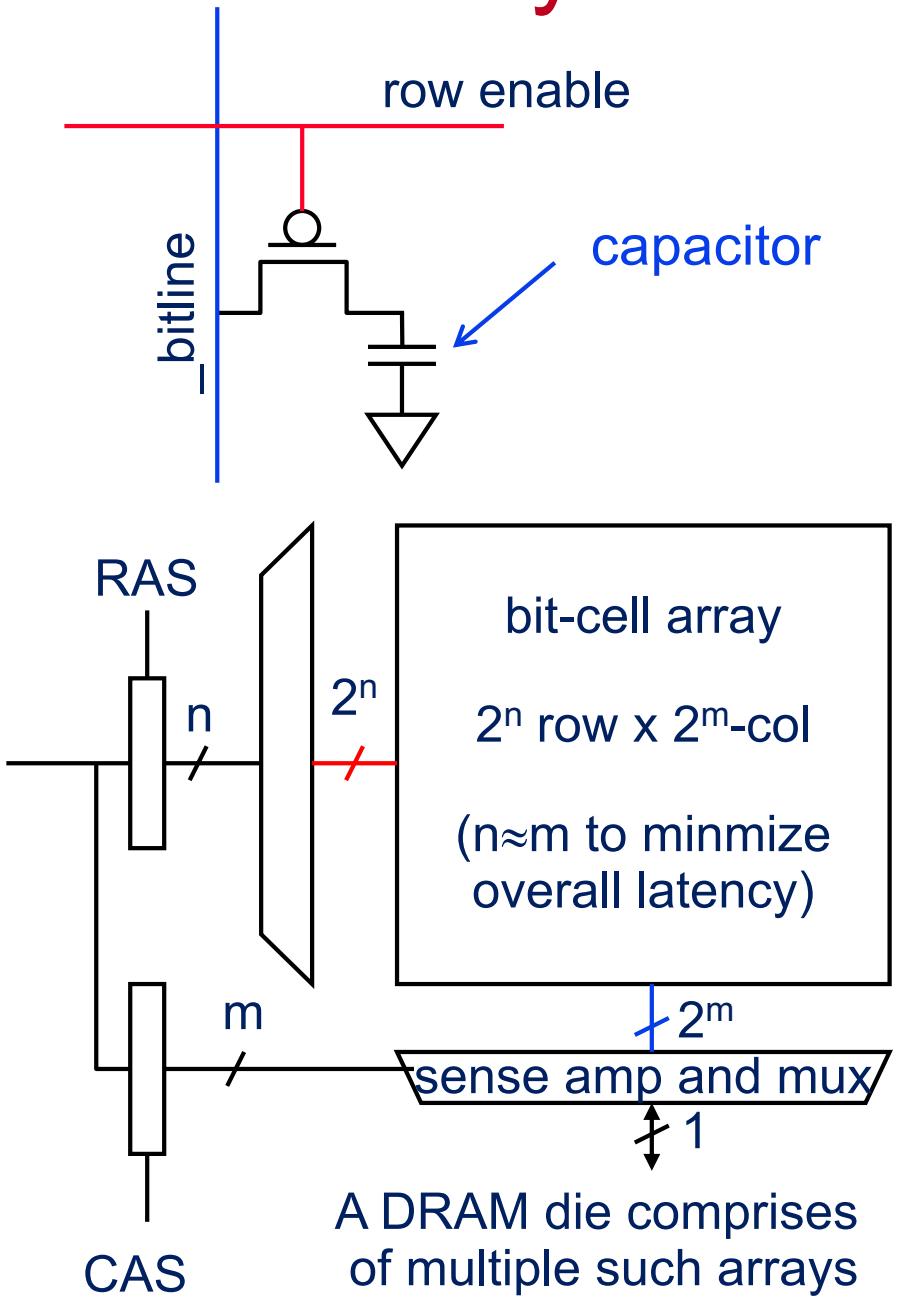
Write: Drive Bit Line, Assert Word Line

Need for Refresh Cycles: storage decay in ms

Destructive Read-Out

Internal circuits read word and write back

Dynamic RAM (DRAM)



Bits stored as charge in capacitor

- Precharge bitline to Vdd/2
  - Bit cell loses charge when read
    → Bit line charge increases
  - Bit cell drains over time
    → Bit line charge decreases

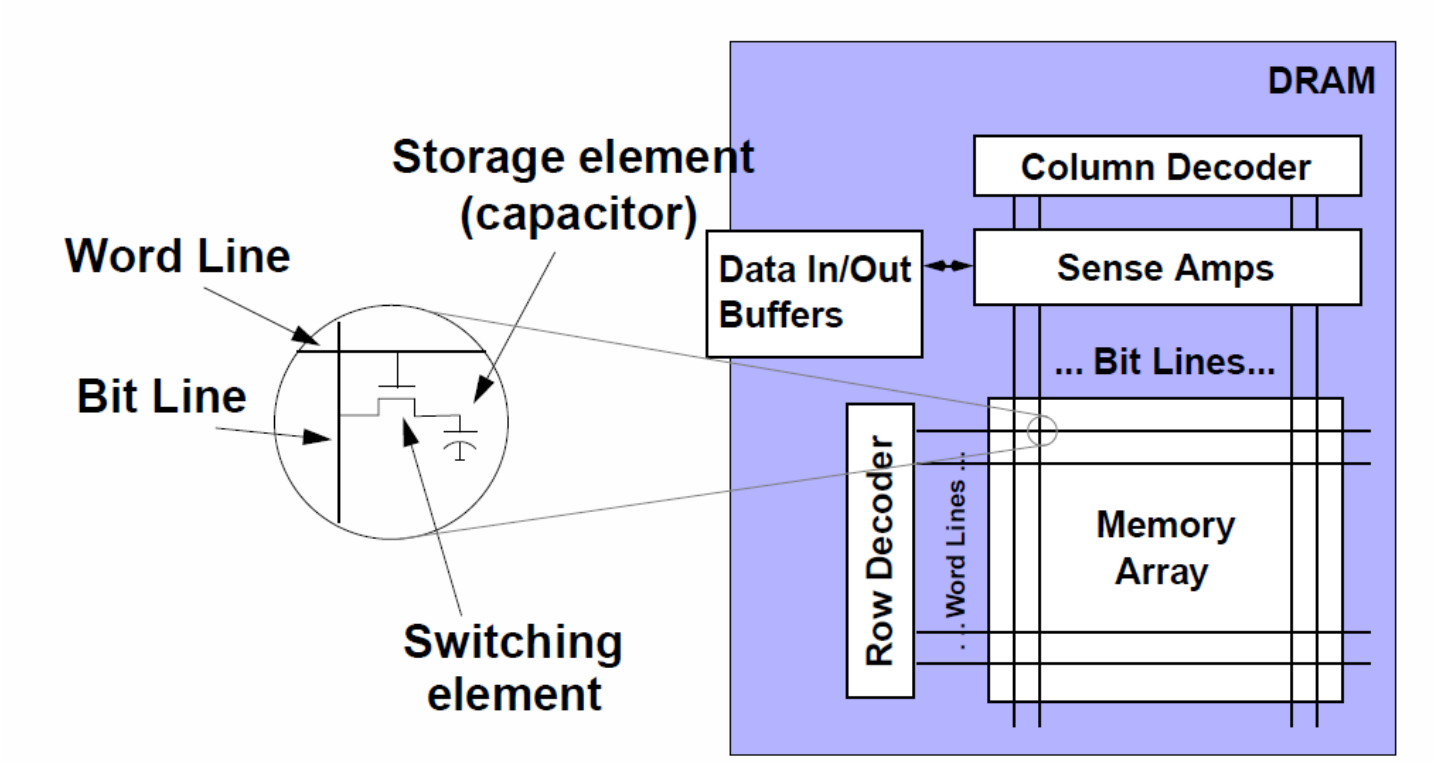
* Disruptive Read!!

Requires periodic refresh

- Performed by mem controller
- Refresh interval 10's of ms
- DRAM unavailable during refresh

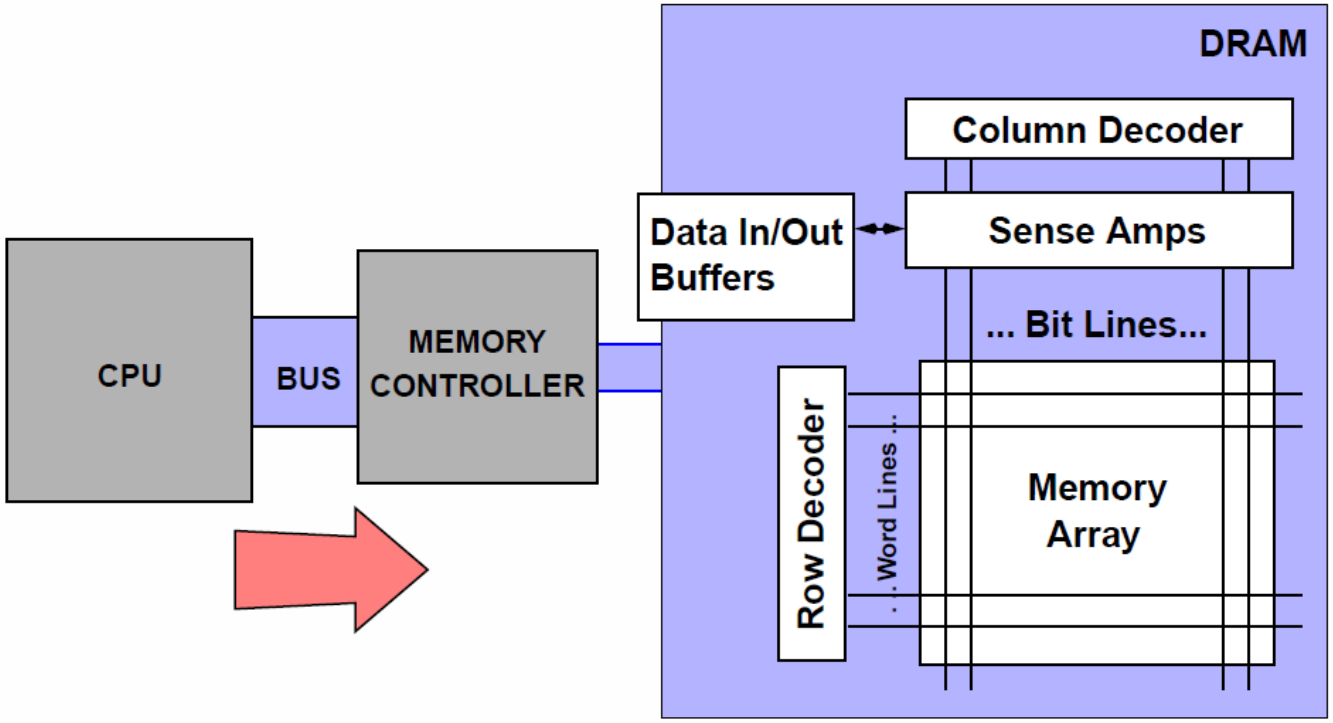
Basics: DRAM organization

- One bit = One Capacitor
- Accessed by Word Line (row) d Bit Line (column)



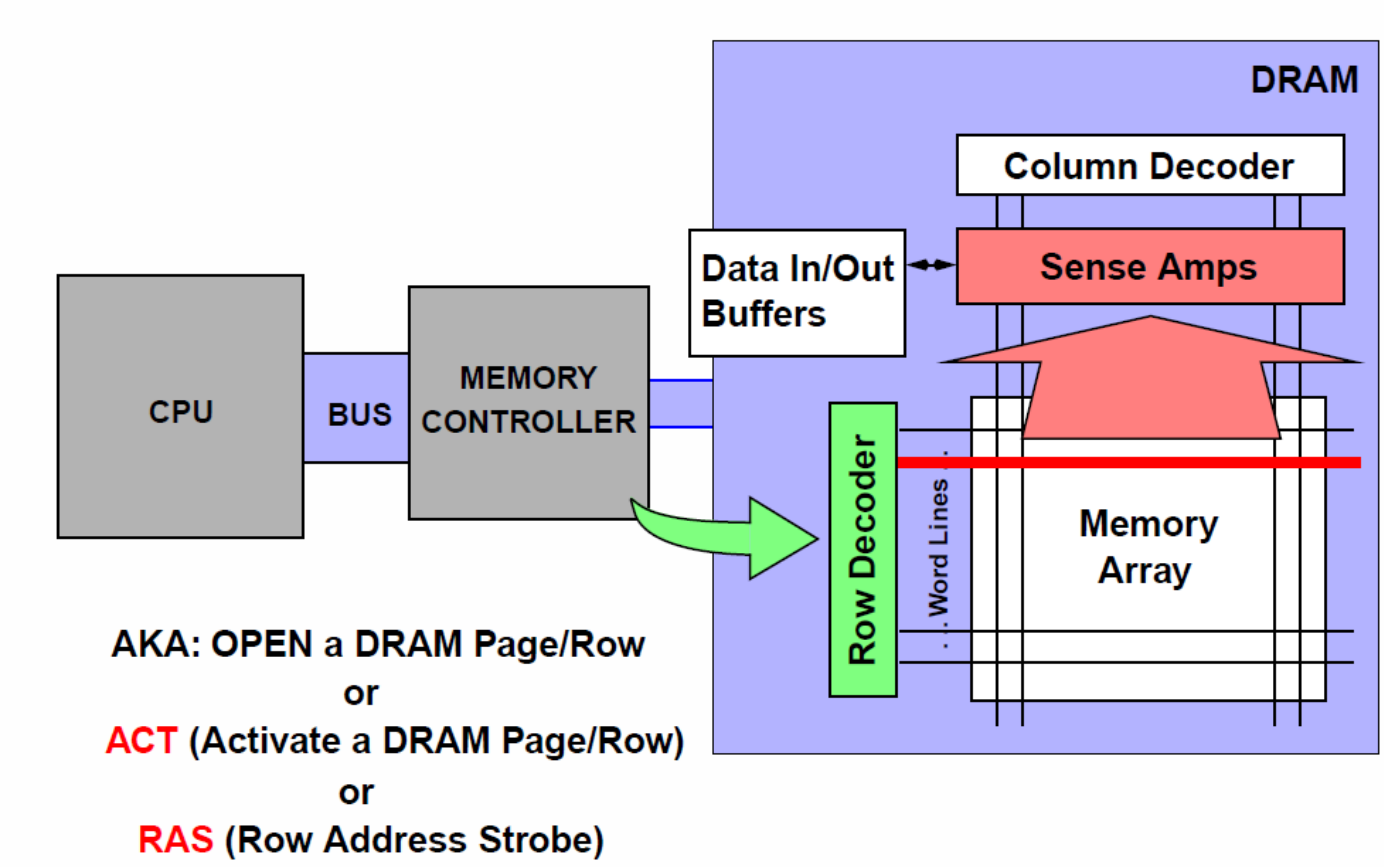
Basics: Access Process (1/5)

(1) BUS TRANSMISSION



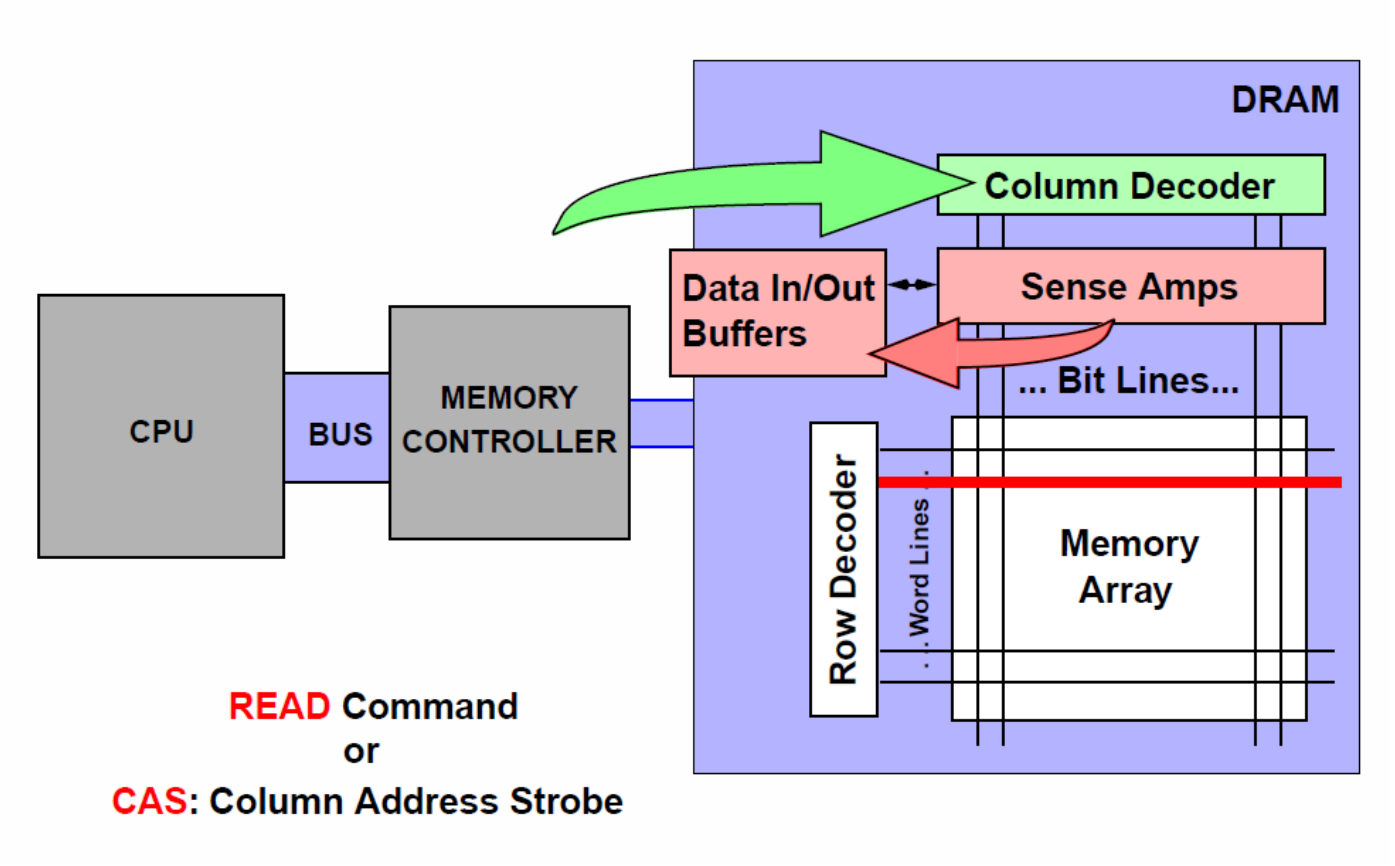
Basics: Access Process (2/5)

(2) (optional PRECHARGE) & ROW ACCESS



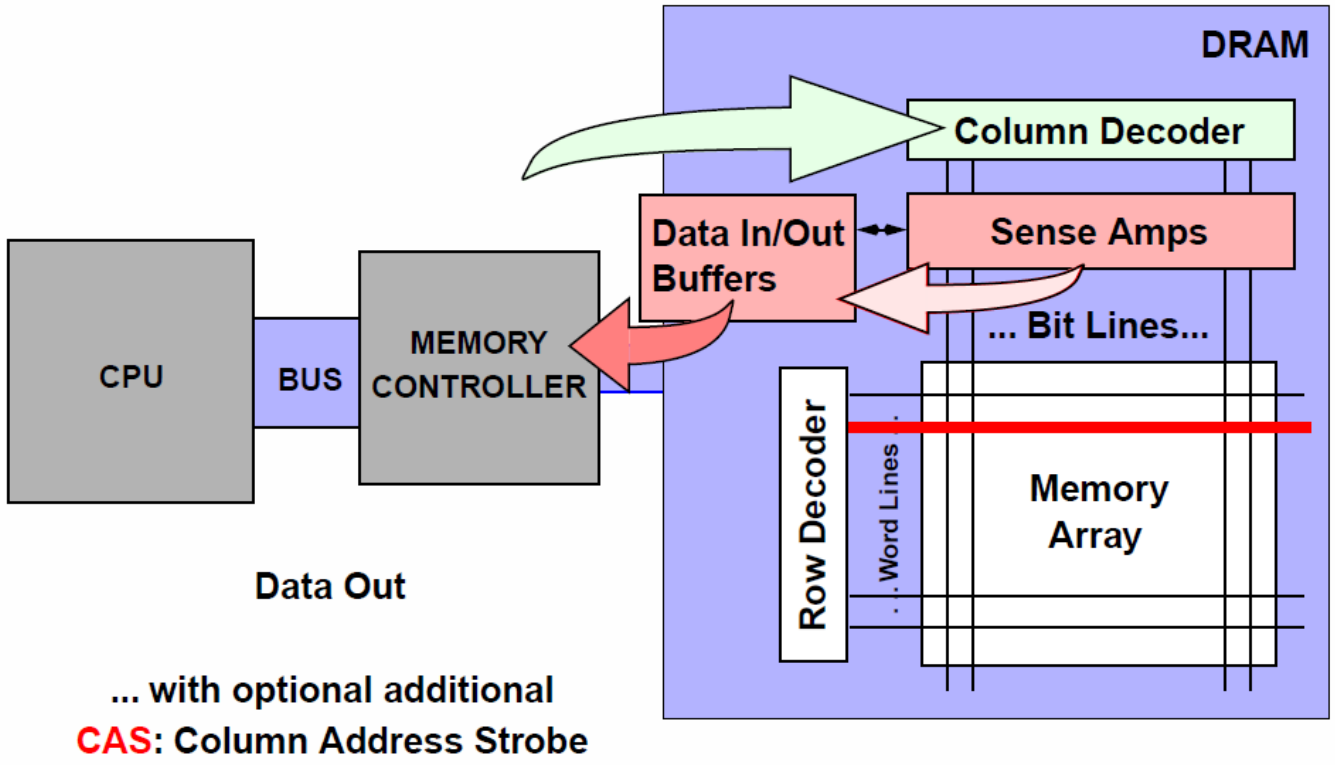
Basics: Access Process (3/5)

(3) COLUMN ACCESS



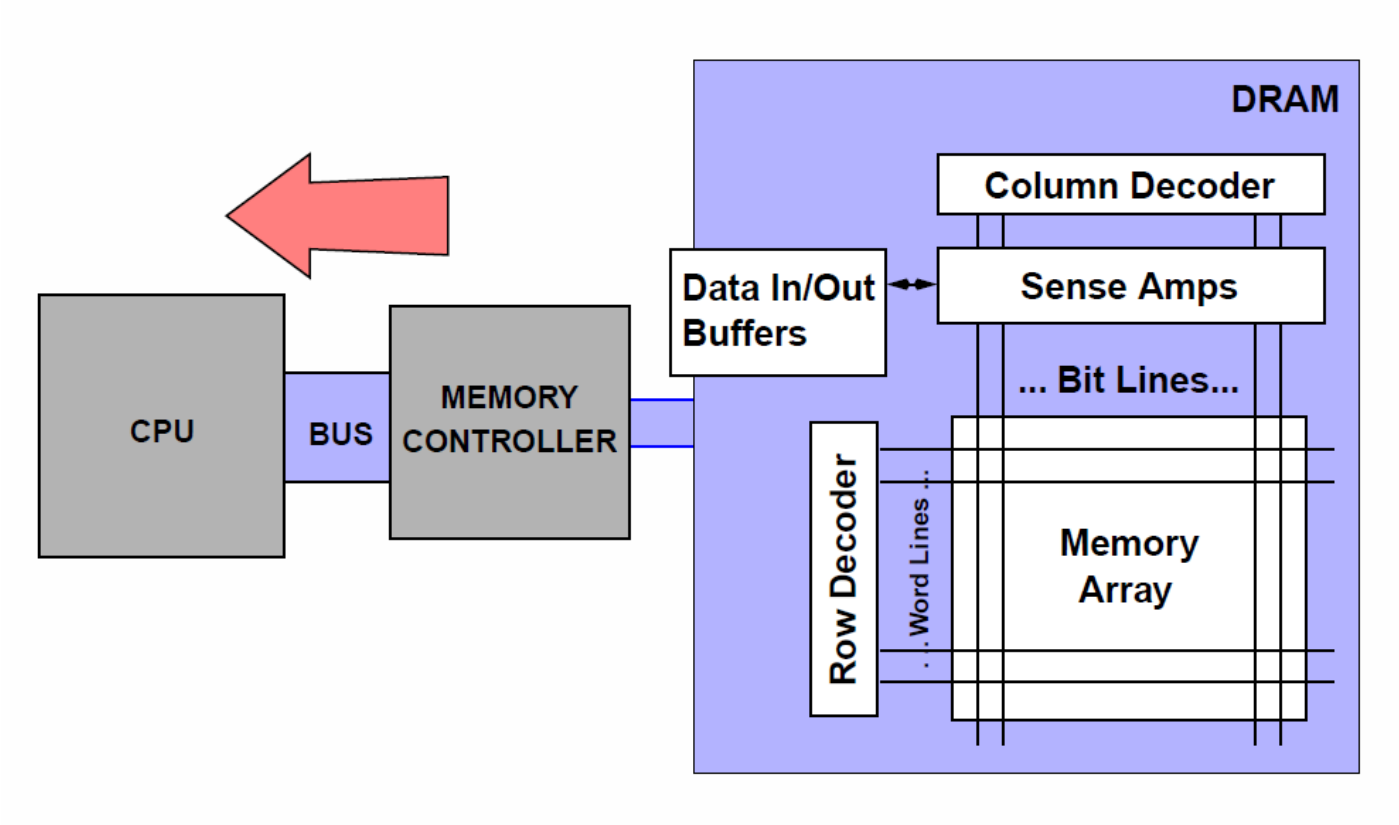
Basics: Access Process (4/5)

(4) DATA TRANSFER

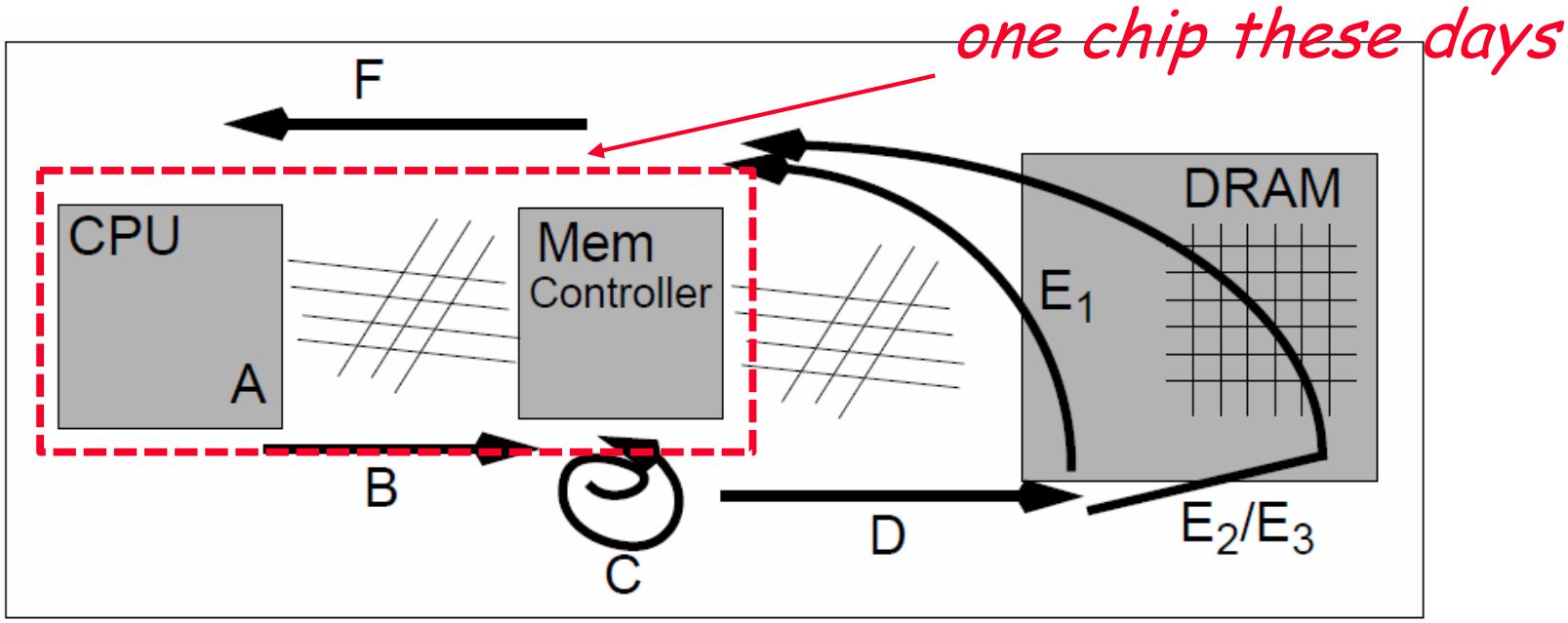


Basics: Access Process (5/5)

(5) BUS TRANSMISSION



DRAM Latency



- DRAM Latency = A + B + C + D + E + F
  - A: Transaction request may be delayed in queue
  - B: Transaction request sent to memory controller
  - C: Transaction request converted as DRAM command sequences
  - D: Commands sent to DRAM
  - **E: Access DRAM : E1 (CAS) or E2 (RAS+CAS) or E3 (PRE+RAS+CAS)**
  - F: Transaction sent back to CPU

Dynamic RAMs

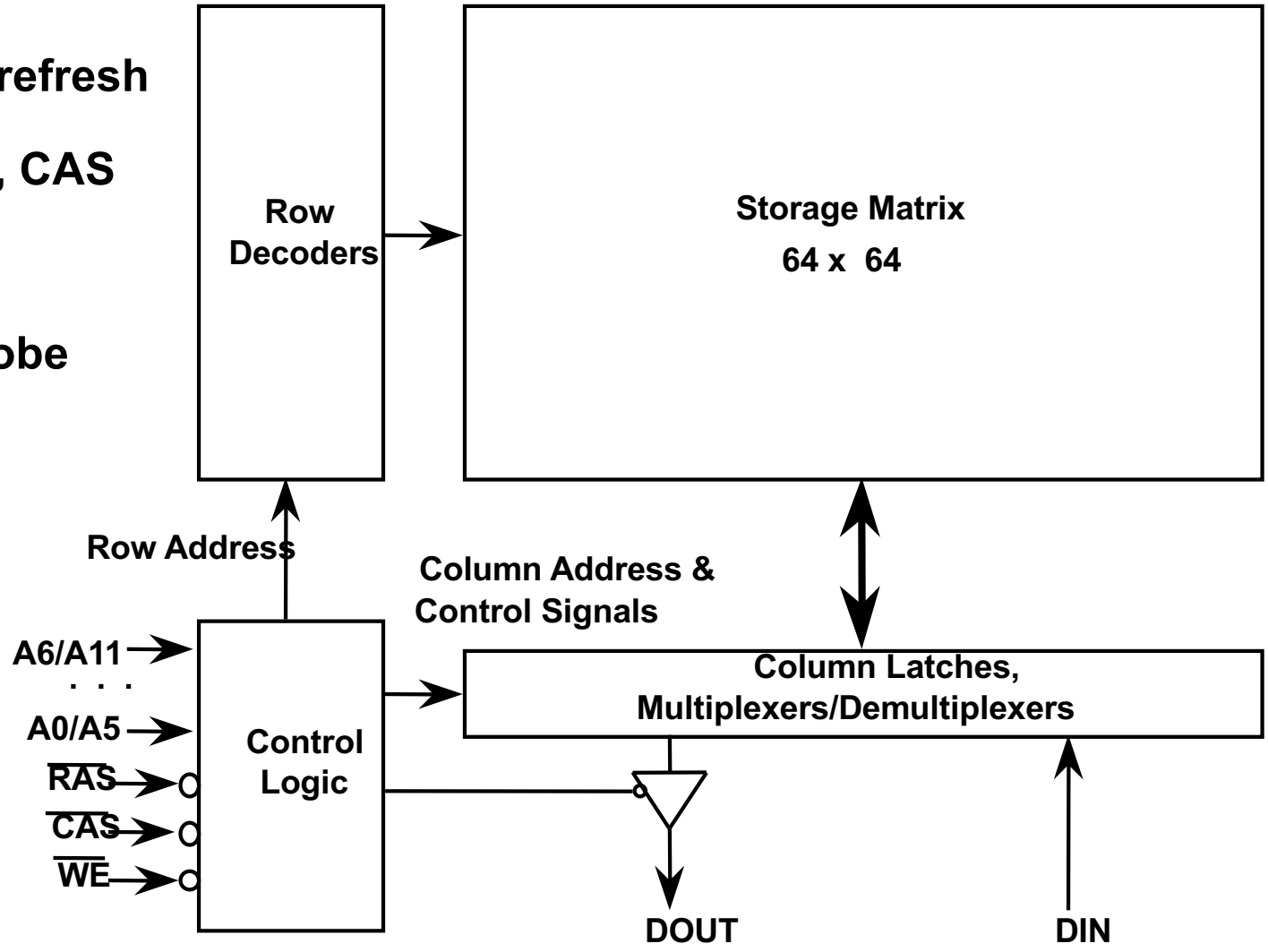
DRAM Organization

Long rows to simplify refresh

Two new signals: RAS, CAS

- Row Address Strobe
- Column Address Strobe

Replace Chip Select

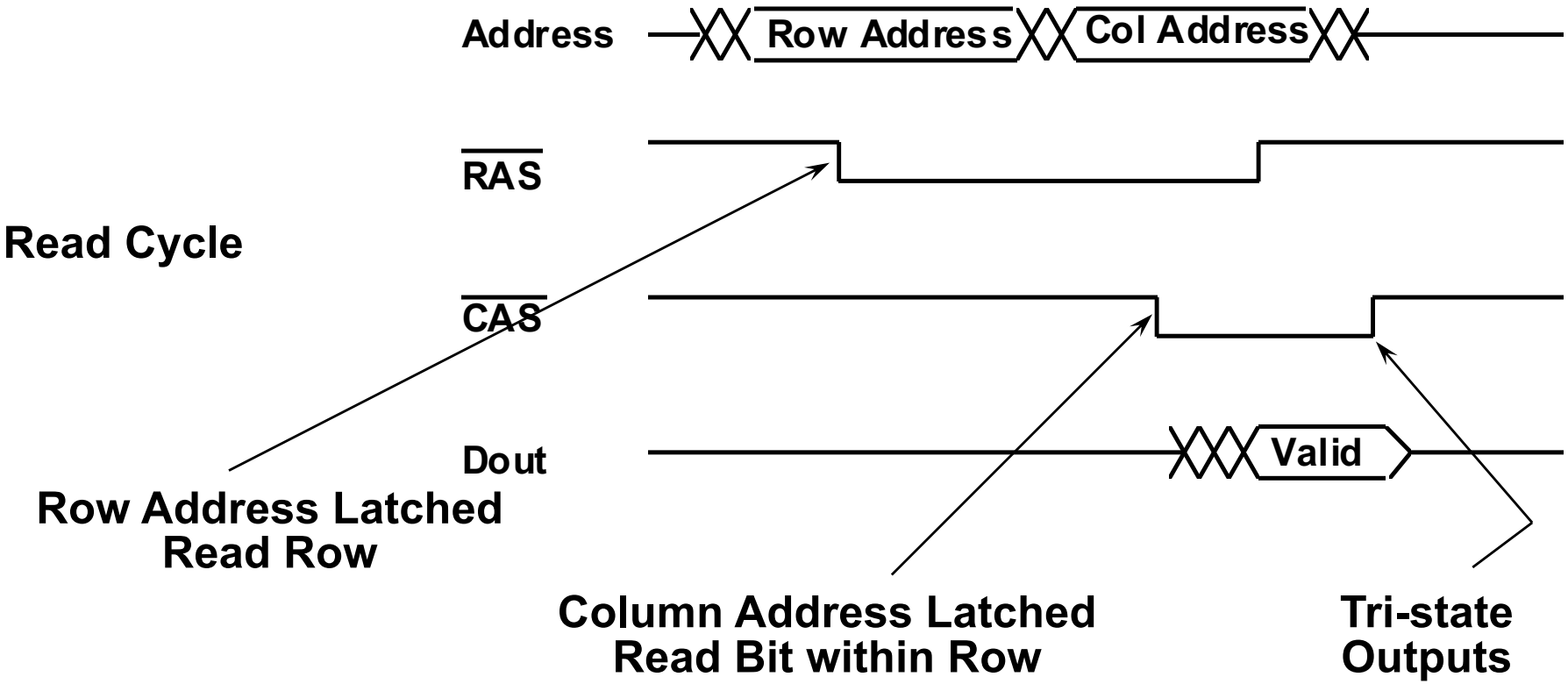


Dynamic RAMs

RAS, CAS Addressing

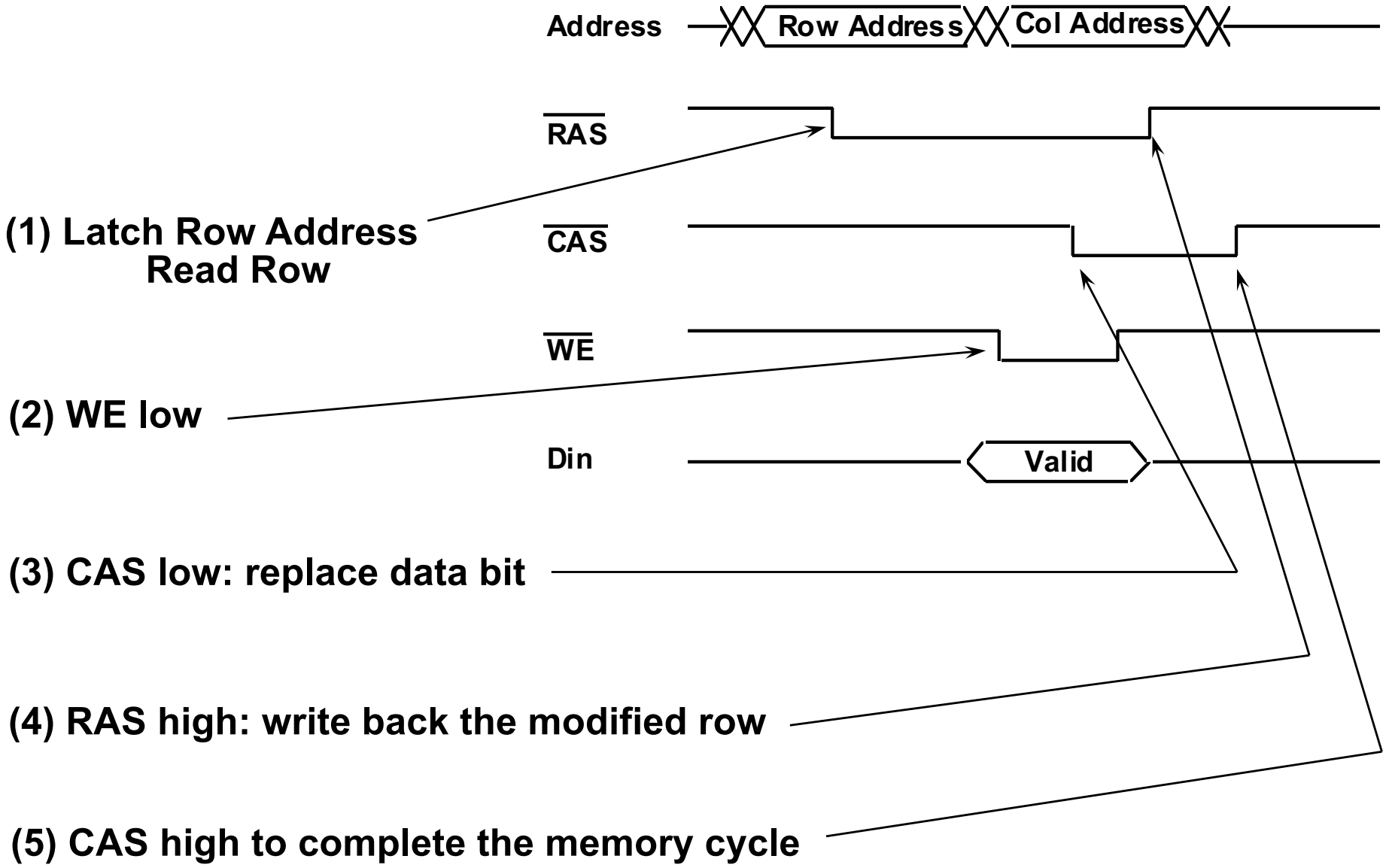
Even to read 1 bit, an entire 64-bit row is read!

Separate addressing into two cycles: Row Address, Column Address
Saves on package pins, speeds RAM access for sequential bits!



Dynamic RAMs

Write Cycle Timing



Dynamic RAMs

RAM Refresh

Refresh Frequency:

4096 word RAM -- refresh each word once every 4 ms
→ This is one refresh cycle every 976 ns

Assume 120ns memory access cycle
→ 1 in 8 DRAM accesses

But, RAM can be organized into 64 rows (with a row refresh)
→ This is one refresh cycle every 62.5 μs
→ 1 in 500 DRAM accesses

Large capacity DRAMs have 256 rows, refresh once every 16 μs

RAS-only Refresh (RAS cycling, no CAS cycling)

External controller remembers last refreshed row

"CAS before RAS" signaling

If CAS goes low before RAS, then do refresh

Dynamic RAMs

- DRAM Variations

Page Mode DRAM:

Can read/write a bit within the last accessed row without RAS cycle

RAS, CAS, CAS, . . ., CAS, RAS, CAS, ...

New column address for each CAS cycle

Static Column DRAM:

Similar to page mode, but address bit changes trigger column reads without using multiple CAS cycling

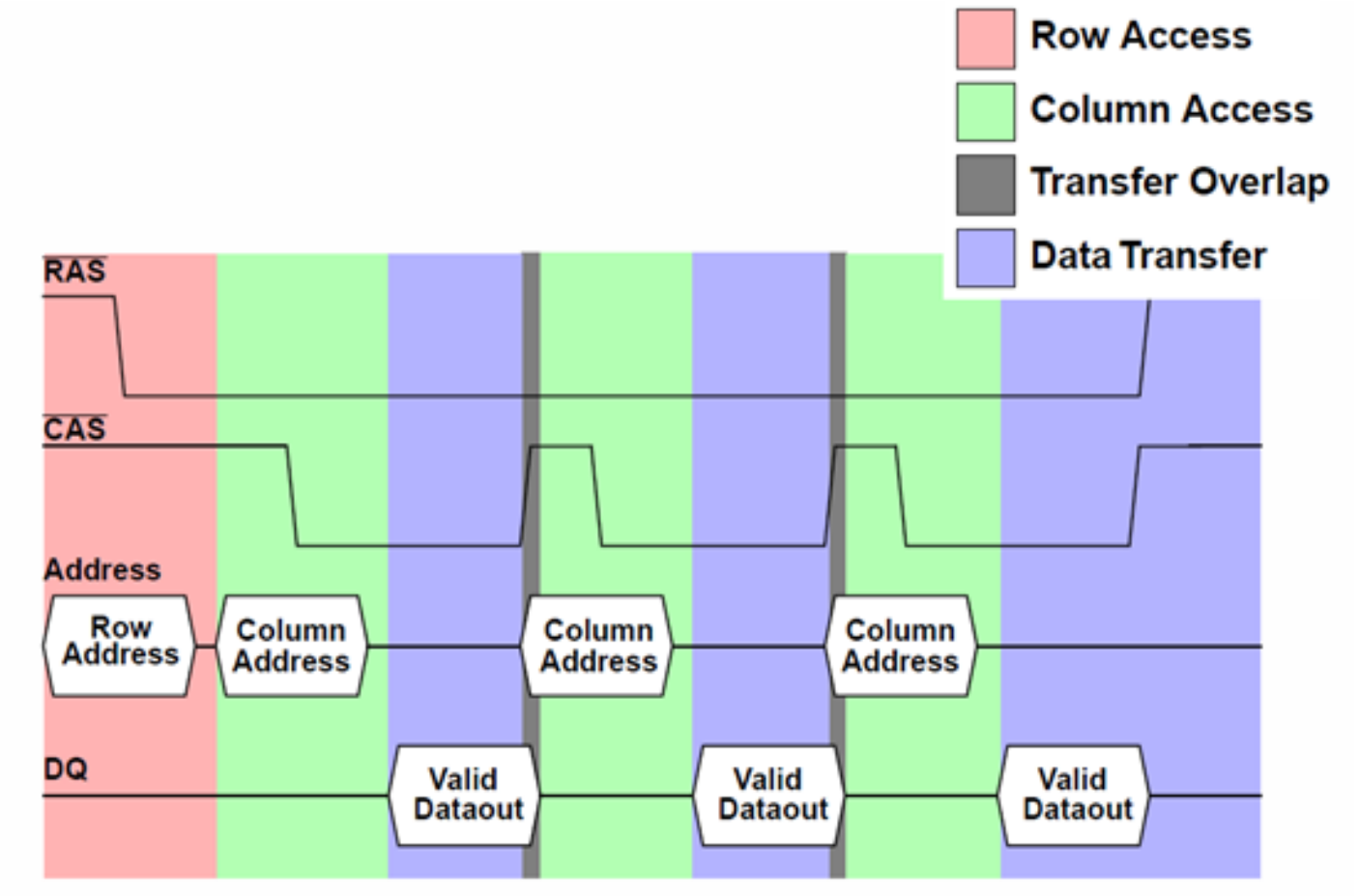
On writes, de-select chip or disable CAS while address changes

Nibble Mode DRAM:

Another variation of page mode.
CAS cycling implies “next column address” in sequence
-- no need to specify column address after first CAS

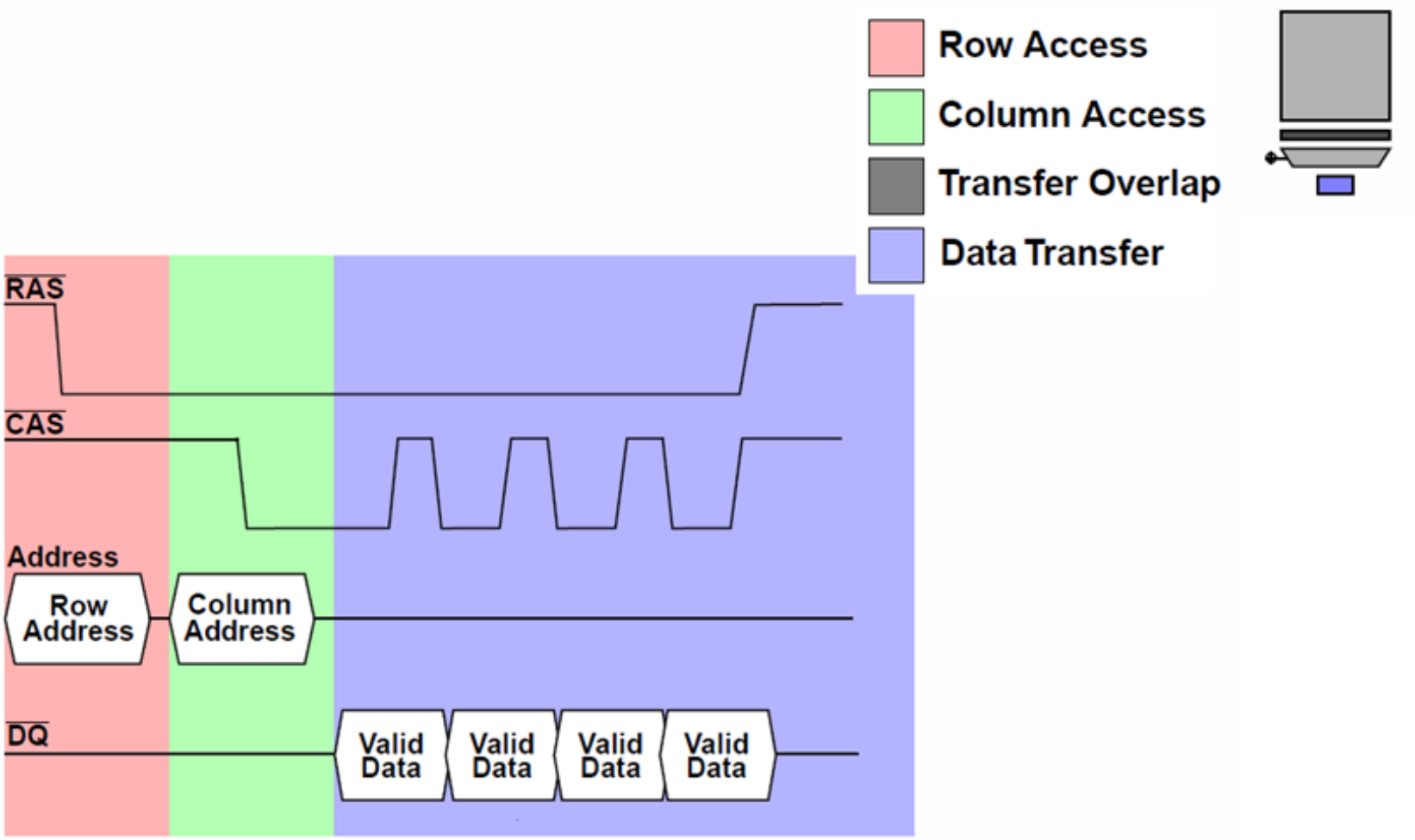
Works for 4 bits at a time (hence "nibble")
RAS, CAS, CAS, CAS, CAS, RAS, CAS, CAS, CAS, CAS, . . .

Read Timing of Fast Page Mode



- One row address, multiple column addresses
- Exploits page buffering (or **Row Buffer**) that DRAMs already require internally

Read Timing of Burst (Nibble) Mode

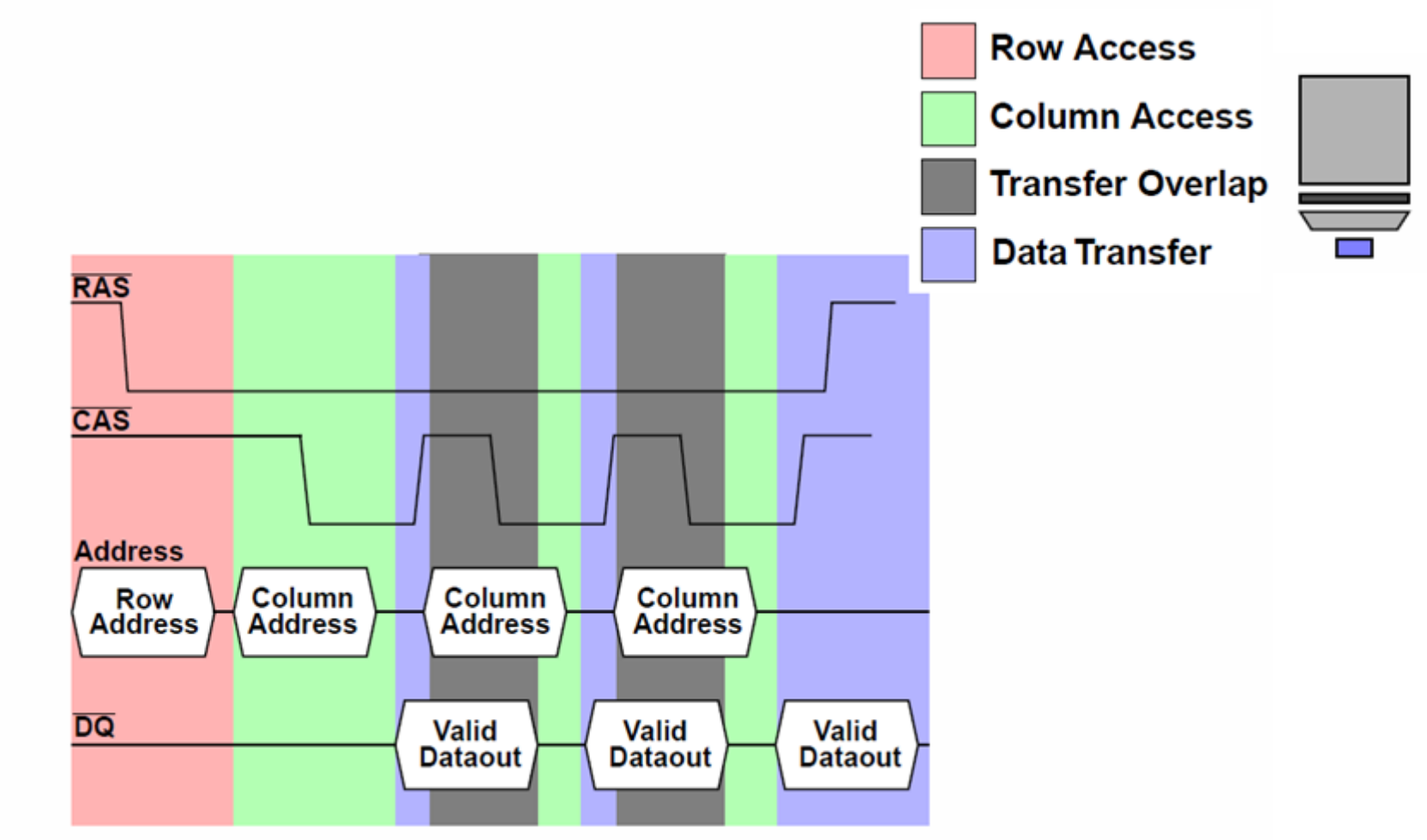


CAS Strobe indicates proceed to next sequential column address

Dynamic RAMs

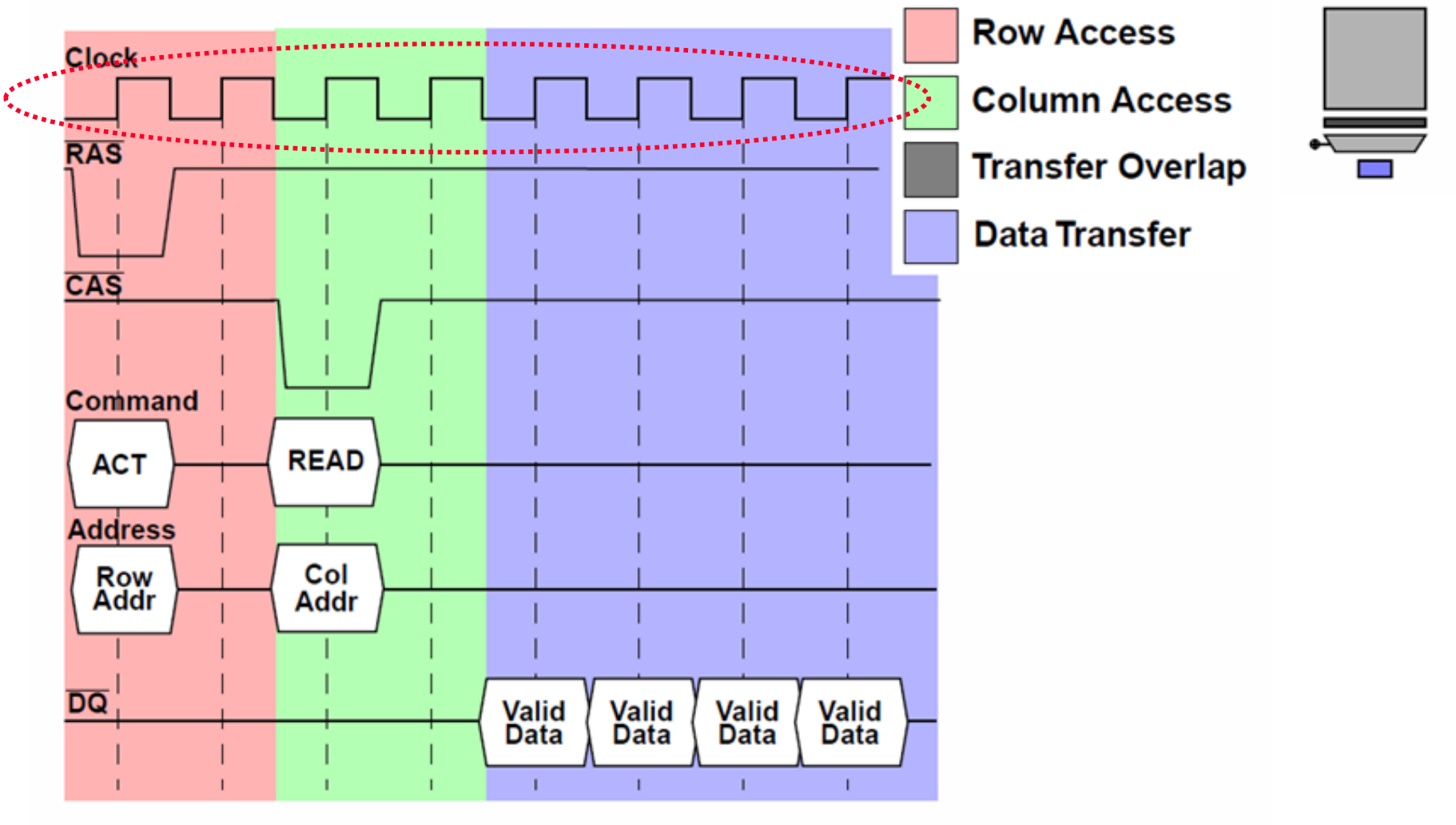
- More advanced DRAM technologies
  - **EDO DRAM** (Extended Data Out DRAM)
    - One access to the memory can begin before the last one has finished
  - **SDRAM** (Synchronous DRAM)
    - Tied to the system clock and is designed to be able to read or write memory in burst mode (after the initial read or write latency) at 1 clock cycle per access (zero wait states)
  - **DDR SDRAM** (Double Data Rate SDRAM)
    - Doubles the bandwidth of the memory by transferring data twice per cycle on both the rising and falling edges of the clock signal
    - DDR2, DDR3, ...
  - **DRDRAM** (Direct Rambus DRAM)
  - **SLDRAM** (Synchronous-Link DRAM)

Read Timing of EDO Mode



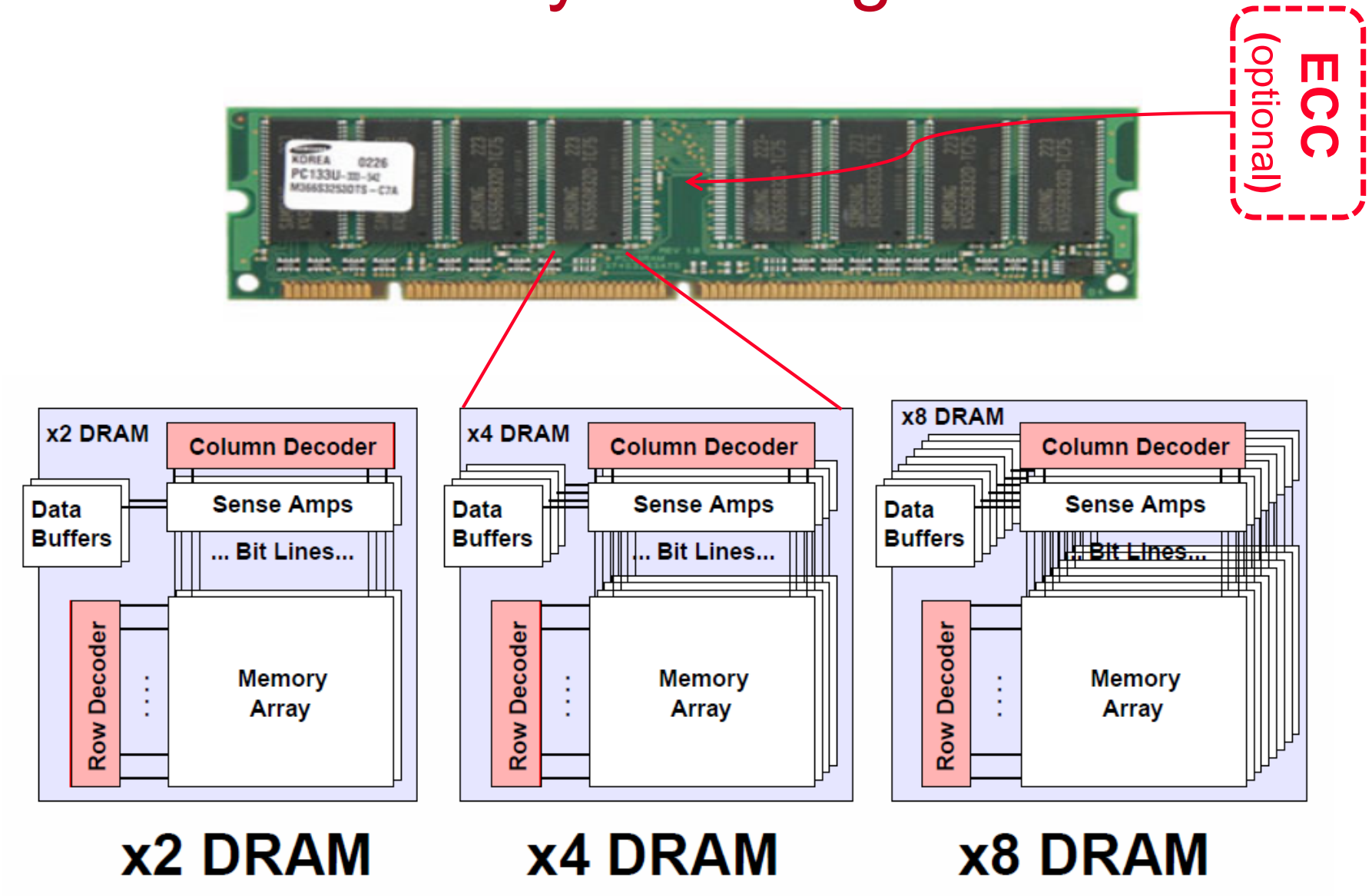
- As in FPM
- But overlapped Column Address assert with Data Out

Synchronous DRAM (SDRAM)



Adding a clock allows faster transfers

Basics: Physical Organization



This is per bank. Typical DRAMs have 2+ banks

SRAM vs. DRAM

- SRAM
  - Faster access using bit-differential sense
    - Pre-charging required
  - Strong storage → doesn't need refreshing
  - But, static power consumption due to Vdd to ground paths

- DRAM
  - Smaller cells → higher density
    - Refreshing required
  - But, dynamic power consumption due to refreshing

FLASH Solid-State Storage

FLASH Cell

Electrically modifiable, non-volatile storage

First developed by Toshiba around 1980

- Evolved from ROM and EEPROM devices

Principle of operation

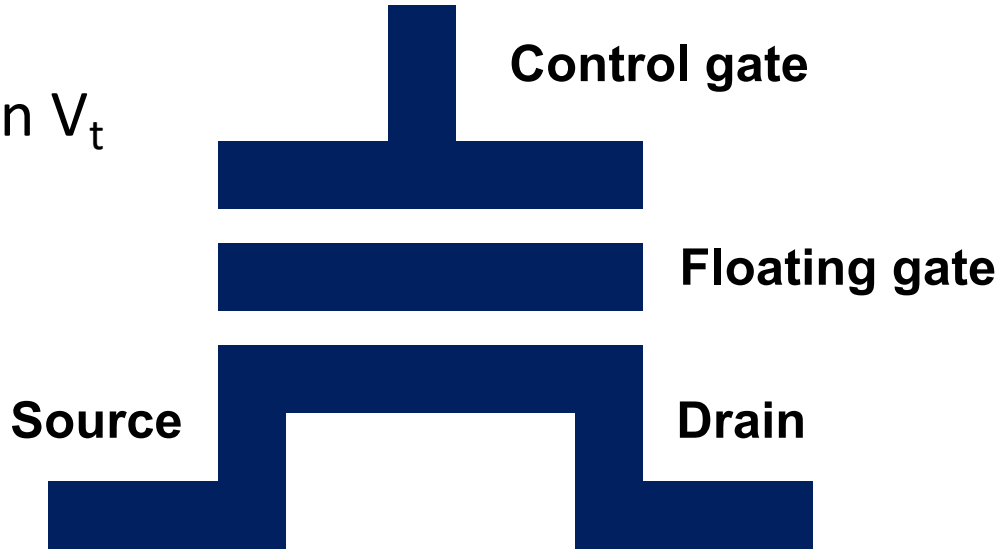
- Transistor with a second "floating" gate
- Floating gate can trap electrons
- Results in detectable change in V_t

  V_t = threshold voltage

  = transistor turn-on point

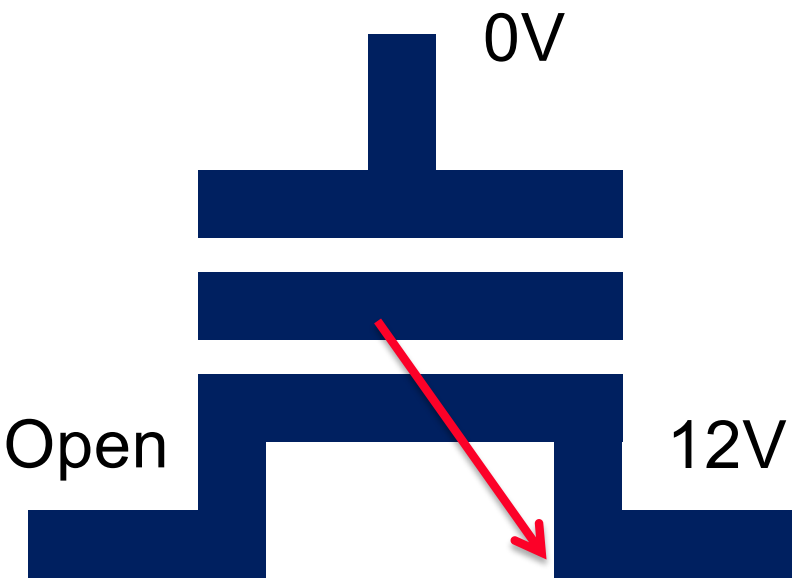
  → High V_t → slow transistor

  Low V_t → fast transistor



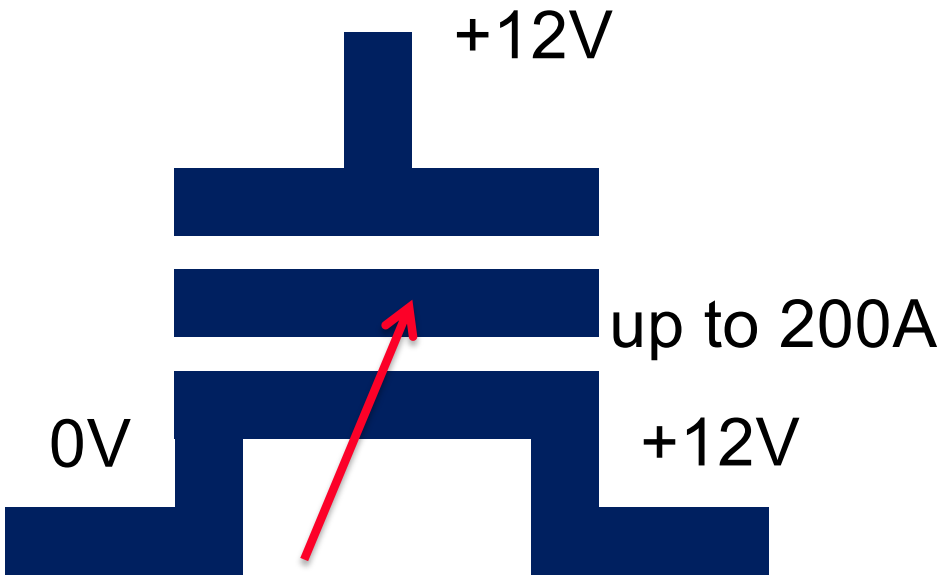
FLASH cell operation (case of NAND)

Erasing to logical "1"



"Quantum tunneling"
Drains charge from FG

Programming (=writing) to logical "0"

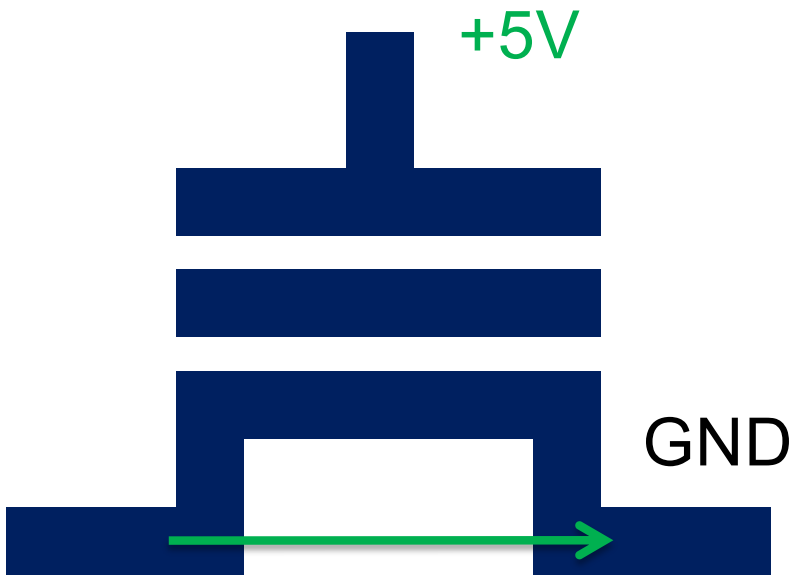


"Hot-electron injection"
traps charge in FG

FLASH cell operation (NAND)

Reading

The strength of flow current between source and gate tells us its logical value : "1" or "0"



Turn on low Vt or High Vt?

Detect I_{on} to read 0 or 1

Value 1:

erased
=less resistant
=fast turn on

Value 0:

programmed
= strong resistance
(due to FTL electrons)
= slow turn on

Question?

Announcements

Reading: Finish reading Chapter 10!

Handouts