

Memory Technology

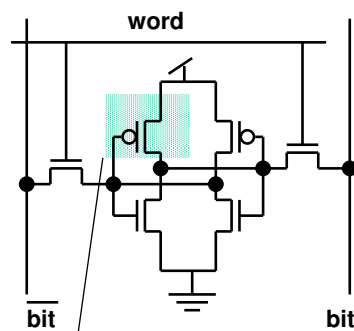
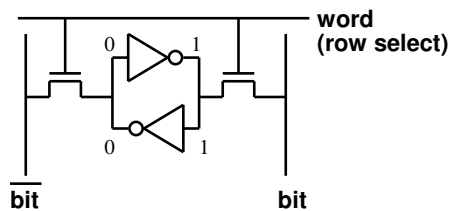
Lecturer: Dr. Hui Annie Guo
 huig@cse.unsw.edu.au
 K17-501F (ext. 57136)

Memory Hierarchy Technology

- **Random Access:**
 - Access time is same for all (random) locations
 - **DRAM: Dynamic Random Access Memory**
 - High density, low power, cheap, slow
 - Dynamic: need to be “refreshed” regularly
 - **SRAM: Static Random Access Memory**
 - Low density, high power, expensive, fast
 - Static: content will last “forever”(until lose power)
- **“Not-so-random” Access Technology:**
 - Access time varies from location to location and from time to time
 - Examples: Disk, CDROM
 - **Sequential Access Technology: access time linear in location (e.g.,Tape)**

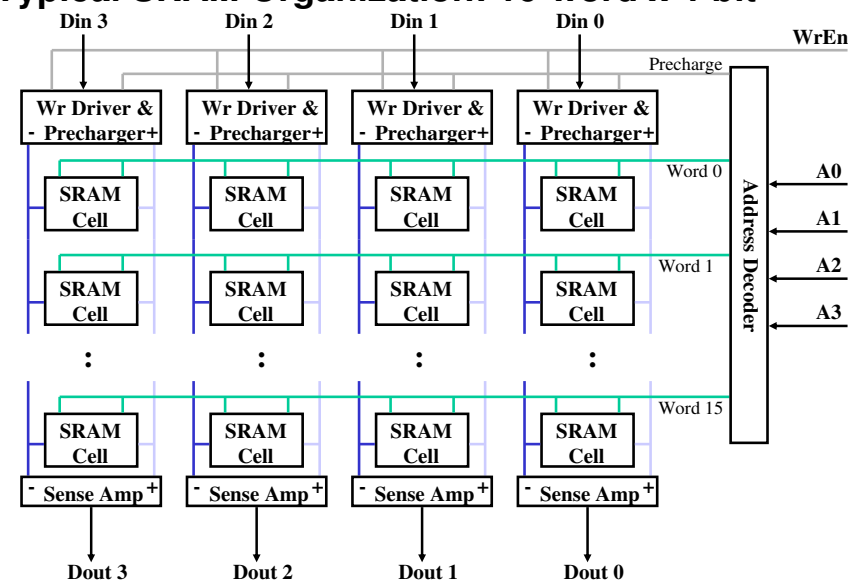
Static RAM Cell

6-Transistor SRAM Cell



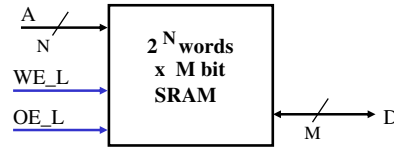
- **Write:**
 1. Drive bit lines (bit=1, bit-bar=0)
 2. Select row
- **Read:**
 1. Precharge bit and bit-bar
 2. Select row
 3. Cell pulls one line low
 4. Sense amp on column detects difference between bit and bit-bar

Typical SRAM Organization: 16-word x 4-bit



Read occurs by default whenever a change in address is sensed

A Typical SRAM



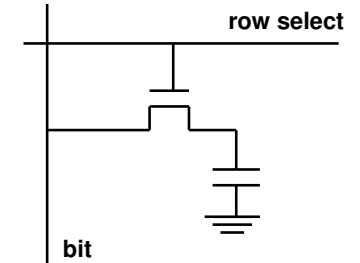
- **Write Enable is usually active low (WE_L)**
- **Din and Dout are combined to save pins:**
 - A new control signal, output enable (OE_L) is needed
 - When WE_L is asserted (Low), OE_L is deasserted (High)
 - D serves as the data input pin
 - When WE_L is deasserted (High), OE_L is asserted (Low)
 - D is the data output pin
 - Both WE_L and OE_L are asserted:
 - Result is unknown. Don't do that!!!

COMP3211/9211

2011S1 wk8_1 P5

1-Transistor Memory Cell (DRAM)

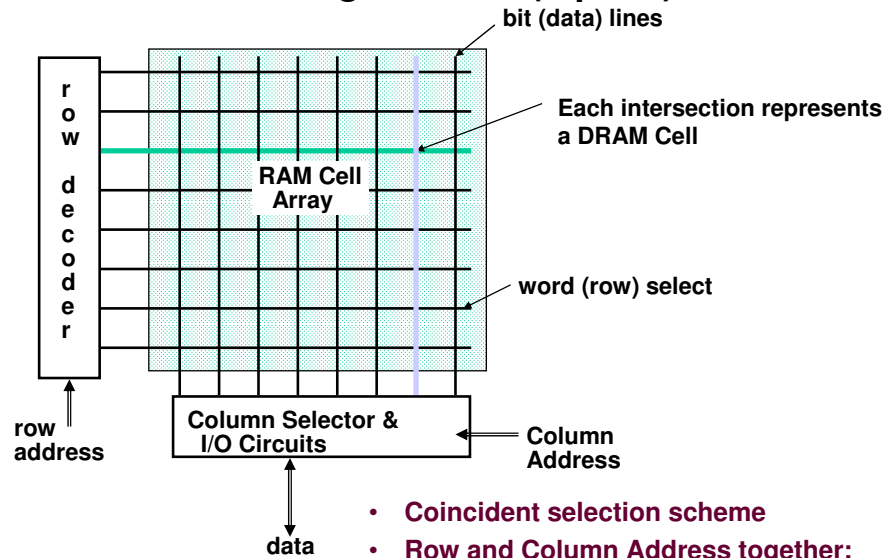
- **Write:**
 1. Drive bit line
 2. Select row
- **Read:**
 1. Precharge bit line
 2. Select row
 3. Cell and bit line share charges
 - Very small voltage changes on the bit line
 4. Sense (sense amp)
 - Can detect changes of ~1 million electrons
 5. Write: restore the value
- **Refresh**
 - Just do a dummy read to every cell.



COMP3211/9211

2011S1 wk8_1 P6

Classical DRAM Organization (square)



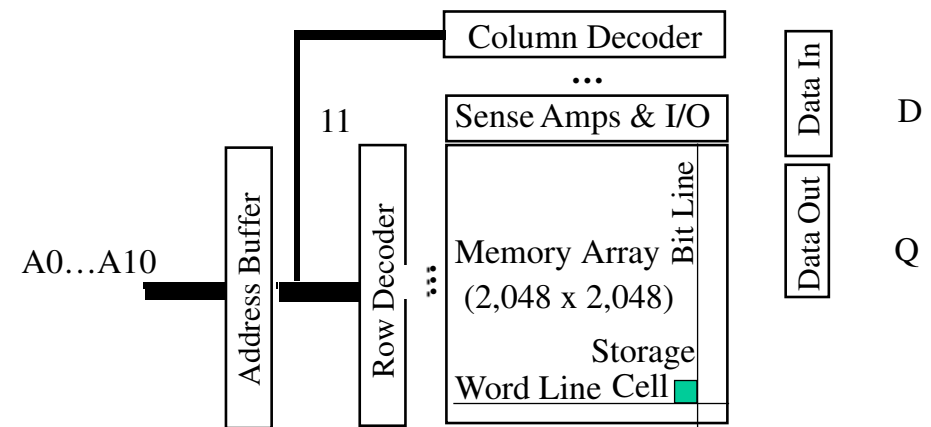
- **Coincident selection scheme**
- **Row and Column Address together:**
 - Select 1 bit a time

COMP3211/9211

2011S1 wk8_1 P7

DRAM logical organization (4 Mbit)

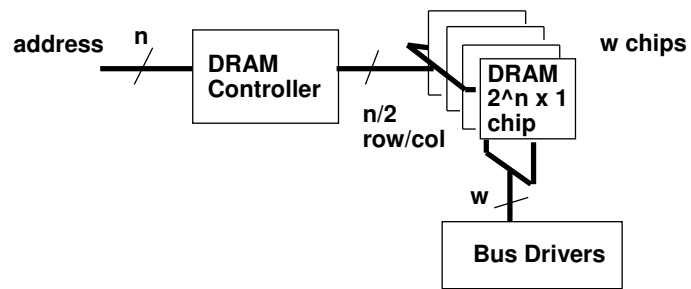
- **Square root of bits per row/column address**



COMP3211/9211

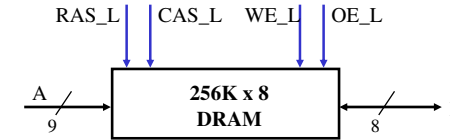
2011S1 wk8_1 P8

Memory Systems



Retrieval = Tcontroller + Taccess + Tdriver

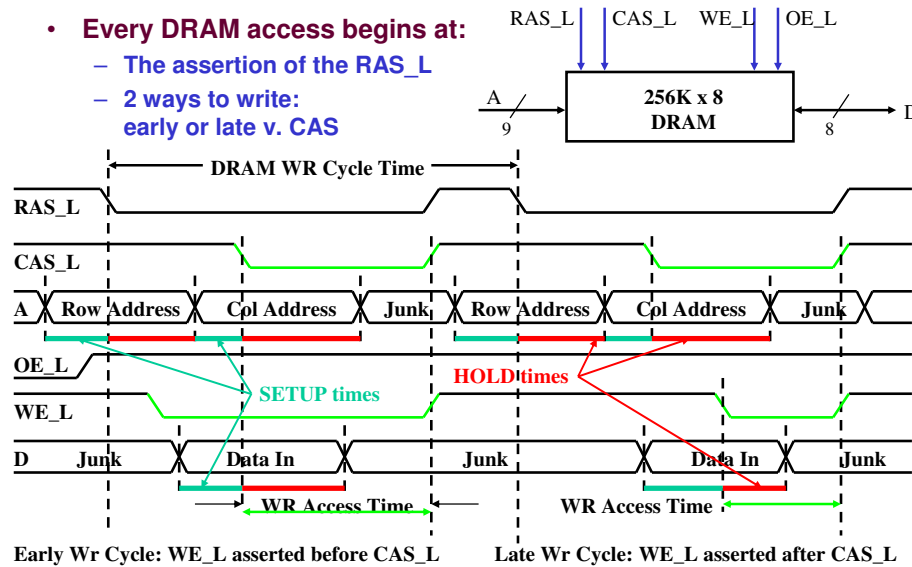
A Typical DRAM



- Control Signals (RAS_L, CAS_L, WE_L, OE_L) are all active low
- Din and Dout are combined (D):
 - When WE_L is asserted (Low), OE_L is deasserted (High)
 - D serves as the data input pin
 - When WE_L is deasserted (High), OE_L is asserted (Low)
 - D is the data output pin
- Row and column addresses share the same pins (A)
 - Controlled using row/column address strobes
 - RAS_L goes low: Pins A are latched in as row address
 - CAS_L goes low: Pins A are latched in as column address
 - RAS/CAS edge-sensitive

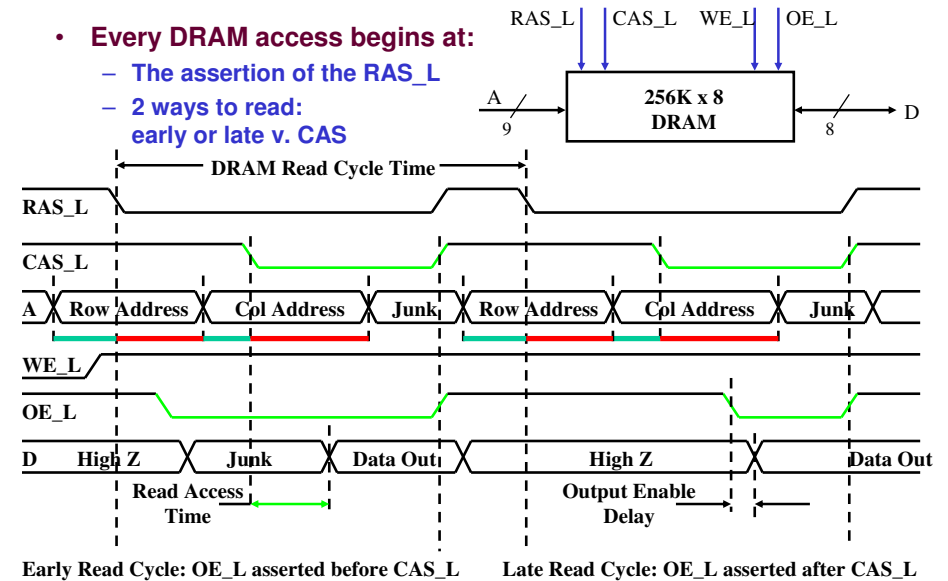
DRAM Write Timing

- Every DRAM access begins at:
 - The assertion of the RAS_L
 - 2 ways to write: early or late v. CAS

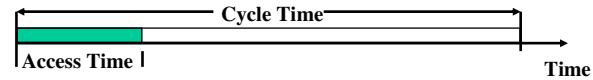


DRAM Read Timing

- Every DRAM access begins at:
 - The assertion of the RAS_L
 - 2 ways to read: early or late v. CAS



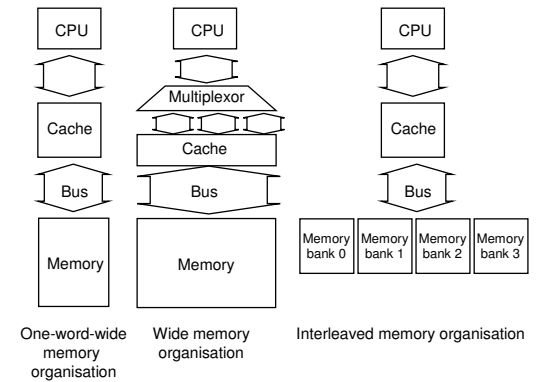
Cycle Time versus Access Time



- **DRAM (Read/Write) Cycle Time >> DRAM (Read/Write) Access Time**
- **DRAM (Read/Write) Cycle Time :**
 - The minimal time between independent accesses, or
 - How frequently we can initiate a new access
 - Limits bandwidth
- **DRAM (Read/Write) Access Time:**
 - The time from receiving the address to delivering the data, or
 - How quickly the data is available?

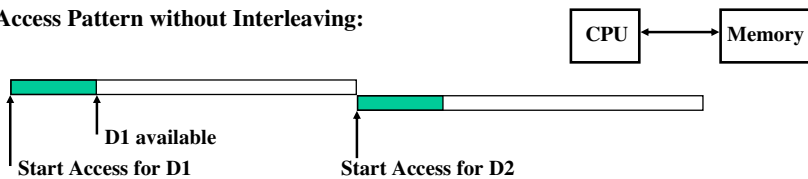
Boosting Main Memory Bandwidth

- **Simple:**
 - CPU, Cache, Bus, Memory same width (32 bits)
- **Wide:**
 - CPU/Mux 1 word; Mux/Cache, Bus, Memory N words (Alpha: 64 bits & 256 bits)
- **Interleaved:**
 - CPU, Cache, Bus 1 word: Memory N Modules (4 Modules); example is *word interleaved*

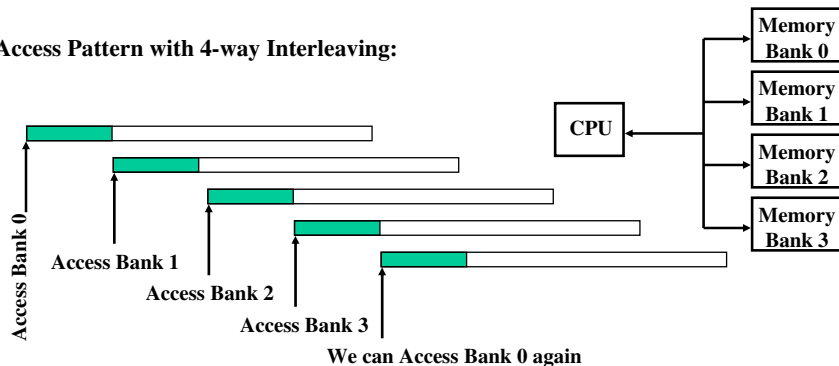


Increasing Bandwidth - Interleaving

Access Pattern without Interleaving:



Access Pattern with 4-way Interleaving:



Main Memory Performance

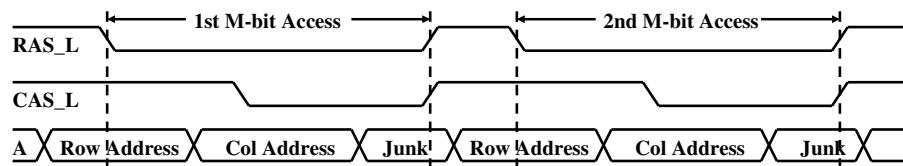
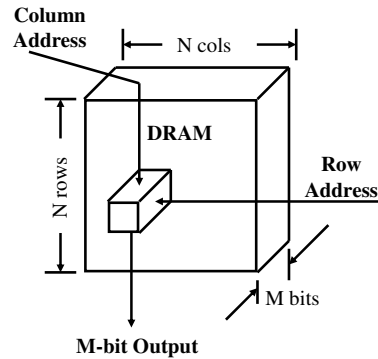
- **Timing model**
 - 1 cycle to send address,
 - 6 cycles access time, 1 cycle to send data
 - Cache Block is 4 words (to get 4 consecutive words)
- **Simple M.** = $4 \times (1+6+1) = 32$ cycles
- **Wide M.** = $1 + 6 + 1 = 8$ cycles
- **Interleaved M.** = $1 + 6 + 4 \times 1 = 11$ cycles

Address	Bank 0	Address	Bank 1	Address	Bank 2	Address	Bank 3
0		1		2		3	
4		5		6		7	
8		9		10		11	
12		13		14		15	

Page Mode DRAM: Motivation

- **Regular DRAM Organization:**

- N rows x N column x M-bit
- Read & Write M-bit at a time
- Each M-bit access requires a RAS / CAS cycle



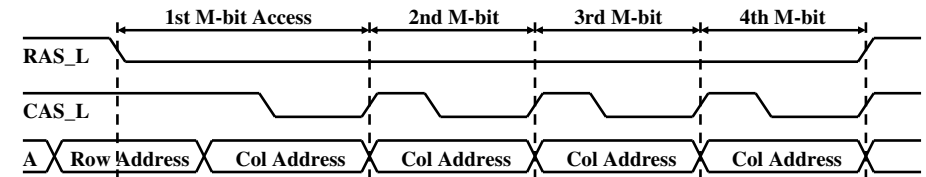
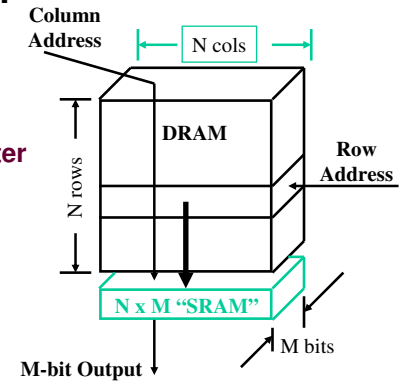
COMP3211/9211

2011S1 wk8_1 P17

Fast Page Mode Operation

- **Fast Page Mode DRAM**

- N x M “SRAM” to save a row
- **After a row is read into the register**
 - Only CAS is needed to access other M-bit blocks on that row
 - RAS_L remains asserted while CAS_L is toggled



COMP3211/9211

2011S1 wk8_1 P18

Summary

- **DRAM is slow but cheap and dense:**
 - Good choice for presenting the user with a BIG memory system
- **SRAM is fast but expensive and not very dense:**
 - Good choice for providing the user FAST access time.
- **There are a number of approaches that can increase the memory bandwidth.**

COMP3211/9211

2011S1 wk8_1 P19