Outline of Today’s Lecture

- Memory Technology: SRAM
- Memory Technology: DRAM
- Page Mode (EDO) DRAM
- RAMBUS
- Synchronous DRAM (SDRAM)
The Big Picture: Where are We Now?

- The Five Classic Components of a Computer

- Today’s Topic: Memory System

Memory Technologies
- Random Access:
  - “Random” is good: access time is the same for all locations
  - DRAM: Dynamic Random Access Memory
    - High density, low power, cheap, slow
    - Dynamic: needs 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)

- The Main Memory: DRAMs

- Caches: SRAMs
Random Access Memory (RAM) Technology

- Why do computer professionals need to know about RAM technology?
  - Processor performance is usually limited by memory latency and bandwidth.
  - Latency: The time it takes to access a word in memory.
  - Bandwidth: The average speed of access to memory (Words/Sec).
  - As IC densities increase, lots of memory will fit on processor chip
    - Tailor on-chip memory to specific needs.
      - Instruction cache
      - Data cache
      - Write buffer

- What makes RAM different from a bunch of flip-flops?
  - Density: RAM is much more denser
  - Speed: RAM access is slower than flip-flop (register) access.

Technology Trends

<table>
<thead>
<tr>
<th></th>
<th>Capacity</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic</td>
<td>2x in 3 years</td>
<td>2x in 3 years</td>
</tr>
<tr>
<td>DRAM</td>
<td>4x in 3 years</td>
<td>1.4x in 10 years</td>
</tr>
<tr>
<td>Disk</td>
<td>2x in 3 years</td>
<td>1.4x in 10 years</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Size</th>
<th>Cycle Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>64 Kb</td>
<td>250 ns</td>
</tr>
<tr>
<td>1983</td>
<td>256 Kb</td>
<td>220 ns</td>
</tr>
<tr>
<td>1986</td>
<td>1 Mb</td>
<td>190 ns</td>
</tr>
<tr>
<td>1989</td>
<td>4 Mb</td>
<td>165 ns</td>
</tr>
<tr>
<td>1992</td>
<td>16 Mb</td>
<td>145 ns</td>
</tr>
<tr>
<td>1995</td>
<td>64 Mb</td>
<td>120 ns</td>
</tr>
<tr>
<td>1998</td>
<td>256 MB</td>
<td>75 ns</td>
</tr>
</tbody>
</table>
Static RAM Cell

6-Transistor SRAM Cell

- Write:
  1. Drive bit lines (bit=1, bit=0)
  2. Select row

- Read:
  1. Precharge bit and bit to Vdd
  2. Select row
  3. Cell pulls one line low
  4. Sense amp on column detects difference between bit and bit

Typical SRAM Organization: 16-word x 4-bit

Din 3  Din 2  Din 1  Din 0  WrEn
Wr Driver & Precharger*  Wr Driver & Precharger*  Wr Driver & Precharger*  Wr Driver & Precharger*
SRAM Cell  SRAM Cell  SRAM Cell  SRAM Cell
SRAM Cell  SRAM Cell  SRAM Cell  SRAM Cell
SRAM Cell  SRAM Cell  SRAM Cell  SRAM Cell
SRAM Cell  SRAM Cell  SRAM Cell  SRAM Cell
Sense Amp*  Sense Amp*  Sense Amp*  Sense Amp*
Dout 3  Dout 2  Dout 1  Dout 0

Address Decoder
A0  A1  A2  A3
Word 0  Word 1  Word 15
Precharge

cps 220 memory
©GK Fall1998
Logic Diagram of 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
  - WE_L is asserted (Low), OE_L is disasserted (High)
    - D serves as the data input pin
  - WE_L is disasserted (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!!!
1-Transistor Memory Cell (DRAM)

- **Write:**
  1. Drive bit line
  2. Select row

- **Read:**
  1. Precharge bit line to Vdd
  2. Select row
  3. Cell and bit line share charges
     1. Very small voltage changes on the bit line
  4. Sense (fancy sense amp)
     1. Can detect changes of ~1 million electrons
  5. Write: restore the value

- **Refresh**
  1. Just do a dummy read to every cell.

Introduction to DRAM

- **Dynamic RAM (DRAM):**
  1. Refresh required
  2. Very high density
  3. Low power (.1 - .5 W active, .25 - 10 mW standby)
  4. Low cost per bit
  5. Pin sensitive (few pins):
     1. Output Enable (OE_L)
     2. Write Enable (WE_L)
     3. Row address strobe (RAS)
     4. Column address strobe (CAS)
Classical DRAM Organization (square)

- Row and Column Address together:
  - Select 1 bit a time

Typical DRAM Organization

- Typical DRAMs: access multiple bits in parallel
  - Example: 2 Mb DRAM = 256K x 8 = 512 rows x 512 cols x 8 bits
  - Row and column addresses are applied to all 8 planes in parallel
Logic Diagram of a Typical DRAM

- Control Signals (RAS_L, CAS_L, WE_L, OE_L) are all active low.
- Din and Dout are combined (D):
  - WE_L is asserted (Low), OE_L is disasserted (High)
    - D serves as the data input pin
  - WE_L is disasserted (High), OE_L is asserted (Low)
    - D is the data output pin
- Row and column addresses share the same pins (A):
  - 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 Timing

- DRAM access is complex. Edge sensitive. Not clocked.
- There are three important periods:
  - Row Access (RAS) >= ~25 ns
  - Column Access (CAS) >= ~25 ns
  - Precharge >= ~25ns
- Page Mode (Extended Data Out):
  - Multiple (random) accesses to the same row.
  - Special features: wrapped burst access for “demand word first”
- Synchronous DRAM:
  - Clocked Access (up to 125 MHz)
  - Multiple banks: can access columns on different open banks with no extra delay.
  - Pipelined access: 3-clocks latency, one “word” per clock throughput. Random access on all open pages with no delay.
  - Burst page mode: can send whole page or portion of a page at a time.
DRAM Timing

- RAMBUS
  - Similar to DRAM: Clocked, Pipelined, Multiple banks, Random access to open pages, etc.
  - Has two clocks:
    - "Slow Clock": (10ns) that defines the global period. All controls are passed using this clock.
    - "Fast Clock": (2.5ns) that is used to pass data on wires. The data is passed on both edges of the clock for a transfer rate of 800MHz. Transfer 8-bit packet at a time (each Slow Clock period).
  - Use 8-bit data bus. Transfers 64 bit every 10ns.

DRAM Write Timing

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

Every Wr Cycle: WE_L asserted before CAS_L
Late Wr Cycle: WE_L asserted after CAS_L
**DRAM Read Timing**

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

The diagram illustrates the DRAM Read Cycle Time, with signals RAS_L, CAS_L, WE_L, and OE_L. The access begins with the assertion of RAS_L and proceeds with the following:

- **Early Read Cycle:** OE_L asserted before CAS_L
- **Late Read Cycle:** OE_L asserted after CAS_L

Every DRAM access begins at:
- The assertion of the RAS_L

2 ways to read: early or late v. CAS

The Read Access Time is measured from the assertion of RAS_L to the output of Data Out. The Output Enable Delay is the time from the assertion of OE_L to the output of Data Out.

The diagram uses red and green arrows to indicate the timing of these events.