Today’s Lecture: DRAM

* DRAM, Xilinx, and You
* DRAM: Bottom-up
* DRAM: Top-down
DDR2 SO-DIMM on ML505 Board

DDR2: Double-Data Rate, 2nd generation

SO-DIMM: Small-Outline, Dual Inline Memory Module

DDR2 SO-DIMM Module

DRAM chips are wired in parallel and run in lockstep.
**Today’s Lecture: DRAM**

**DRAM, Xilinx, and You**

To understand the DRAM controller, you need to understand how a DRAM chip works. Otherwise, it just seems like magic.
Recall: Building a capacitor

- **Top Plate**: Conducts electricity well. (metal, doped polysilicon)
- **Bottom Plate**: Conducts electricity well (metal, doped polysilicon)
- **Dielectric**: An insulator. Does not conduct electricity at all. (air, glass (silicon dioxide))
Recall: Capacitors in action

Because the dielectric is an insulator, and does not conduct.

After circuit “settles” ...

\[ Q = C \cdot V = C \cdot 1.5 \text{ Volts (D cell)} \]

**Q**: Charge stored on capacitor

**C**: The capacitance of the device: function of device shape and type of dielectric.

After battery is removed:

Still, \( Q = C \cdot 1.5 \) Volts

Capacitor “remembers” charge

Storing computational state as charge

State is coded as the amount of energy stored by a device.

State is read by sensing the amount of energy

Problems: noise changes \( Q \) (up or down), parasitics leak or source \( Q \). Fortunately, \( Q \) cannot change instantaneously, but that only gets us in the ballpark.
MOS Transistors

Two diodes and a capacitor in an interesting arrangement. So, we begin with a diode review ...

Diodes in action ...

Resistor | Light emitting diode (LED) | Light on?
---|---|---
| | Yes! |
| | No! |
Diodes: Current vs Voltage

Diode is off
\[ I = -I_0 \]

Diode is on
\[ I = I_0 \exp\left(\frac{V}{V_0}\right) \]

\[ I = I_0 \left[\exp\left(\frac{V}{V_0}\right) - 1\right] \]

Io range: 1 fA to 1 nA  
Vo range: 25 mV to 60 mV

How to make a silicon diode ...

Wafer cross-section

At \( V = 0 \), “hill” too high for electrons to diffuse up.

For holes, going “downhill” is hard.

V controls hill.
Note: IC Diodes are biased “off”!

\[ V_1, V_2 > 0\text{V}. \text{ Diodes “off”, only current is } I_0 \text{ “leakage”}. \]

\[ I = I_0 \exp\left(\frac{V}{V_0}\right) - 1 \]

Anodes of all diodes on wafer connected to ground.

MOS Transistors

Two diodes and a capacitor in an interesting arrangement...
What we want: the perfect switch.

Switch is off. V1 is not connected to V2.

Switch is on. V1 is connected to V2.

We want to turn a p-type region into an n-type region under voltage control.

We need electrons to fill valence holes and add conduction band electrons.

An n-channel MOS transistor (nFET)

Vd = 1V, Vg = 0V, Vs = 0V

Polysilicon gate, dielectric, and substrate form a capacitor.

nFet is off (I is "leakage")

Vd = 1V, Vg = 1V, Vs = 0V

Vg = 1V, small region near the surface turns from p-type to n-type.

nFet is on
Mask set for an n-Fet (circa 1986)

\[ V_d = 1V \quad V_g = 0V \quad V_s = 0V \]

<table>
<thead>
<tr>
<th>#1: n+ diffusion</th>
<th>#2: poly (gate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#3: diff contact</td>
<td>#4: metal</td>
</tr>
</tbody>
</table>

Top-down view:

Modern processes have 6 to 10 metal layers (or more) (in 1986: 2).

Dynamic Memory Cells
Recall: Capacitors in action

Because the dielectric is an insulator, and does not conduct.

\[ I = 0 \]

After circuit “settles” ...

\[ Q = C \cdot V = C \cdot 1.5 \text{ Volts (D cell)} \]

- **Q**: Charge stored on capacitor
- **C**: The capacitance of the device: function of device shape and type of dielectric.

After battery is removed:

Still, \( Q = C \cdot 1.5 \) Volts

Capacitor “remembers” charge

---

**DRAM cell: 1 transistor, 1 capacitor**

“Bit Line” \( \rightarrow \) “Word Line”  \( \leftarrow \) Vdd

Word Line and Vdd run on “z-axis”

---
A 4 x 4 DRAM array (16 bits) ....

Invented after SRAM, by Robert Dennard

United States Patent Office

3,387,286
Patented June 4, 1968

FIELD-EFFECT TRANSISTOR MEMORY
Robert H. Dennard, Croton-on-Hudson, N.Y., assignor to International Business Machines Corporation, Armonk, N.Y., a corporation of New York
Filed July 14, 1967, Ser. No. 653,415
21 Claims, (Cl. 340—173)

prent in disclosing various concepts and structures which have been developed in the application of field-effect transistors to different types of memory applications, the primary thrust up to this time in conventional read-write random access memories has been to connect a plurality of field-effect transistors in each cell in a latch configuration. Memories of this type require a large number of active devices in each cell and therefore each cell re-
DRAM Circuit Challenge #1: Writing

$V_{dd} - V_{th}$. Bad, we store less charge. Why do we not get $V_{dd}$?

$$I_{ds} = k [V_{gs} - V_{th}]^2$$

but "turns off" when $V_{gs} <= V_{th}$!

$V_{gs} = V_{dd} - V_{c}$. When $V_{dd} - V_{c} = V_{th}$, charging effectively stops!

---

DRAM Challenge #2: Destructive Reads

Raising the word line removes the charge from every cell it connects to!

DRAMs write back after each read.
### DRAM Circuit Challenge #3a: Sensing

Assume $C_{cell} = 1 \text{ fF}$

Bit line may have 2000 nFet drains, assume bit line C of $100 \text{ fF}$, or $100 \times C_{cell}$.

$C_{cell}$ holds $Q = C_{cell} \times (V_{dd} - V_{th})$

When we dump this charge onto the bit line, what voltage do we see?

$$dV = \frac{C_{cell} \times (V_{dd} - V_{th})}{100 \times C_{cell}}$$

$$dV = \frac{(V_{dd} - V_{th})}{100} \approx \text{tens of millivolts!}$$

In practice, scale array to get a 60mV signal.

### DRAM Circuit Challenge #3b: Sensing

How do we reliably sense a 60mV signal?

Compare the bit line against the voltage on a “dummy” bit line.

“Dummy” bit line. Cells hold no charge.

Bit line to sense

“sense amp”

Dummy bit line

In practice, scale array to get a 60mV signal.
DRAM Challenge #4: Leakage ...

Parasitic currents leak away charge.

Solution: “Refresh”, by rewriting cells at regular intervals (tens of milliseconds)

DRAM Challenge #5: Cosmic Rays ...

Cell capacitor holds 25,000 electrons (or less). Cosmic rays that constantly bombard us can release the charge!

Solution: Store extra bits to detect and correct random bit flips (ECC).
DRAM Challenge 6: Yield

If one bit is bad, do we throw chip away?

Solution: add extra bit lines (i.e. 80 when you only need 64). During testing, find the bad bit lines, and use high current to burn away “fuses” put on chip to remove them.

Extra bit lines. Used for “sparing”.

Moore’s Law for CPUs and DRAMs

Transistors Per Die

Main driver: device scaling ...

Process Scaling: Why chips don’t fry

Each generation of IC technology, we shrink width and length of cell.

If Ccell and drain capacitances scale together, number of bits per bit line stays constant.

\[
dV \approx 60 \text{ mV} = \frac{[C_{cell} \cdot (V_{dd} - V_{th})]}{100 \cdot C_{cell}}
\]

**Problem 1:** Number of arrays per chip grows!

**Problem 2:** Vdd may need to scale down too!

**Solution:** Constant Innovation of Cell Capacitors!

---

**Poly-diffusion Ccell is ancient history**

“Bit Line”  “Word Line”  Vdd

Word Line  Vdd

“Bit Line”

Word Line and Vdd run on “z-axis”
Early replacement: “Trench” capacitors

Figure 4
SEM photomicrograph of 0.25-μm trench DRAM cell suitable for scaling to 0.15μm and below. Reprinted with permission from [17]; © 1995 IEEE.

Final generation of trench capacitors

The companies that kept scaling trench capacitors for commodity DRAM chips went out of business.
Modern cells: “stacked” capacitors

In the labs: Vertical cell transistors...

A 31 ns Random Cycle VCAT-Based 4F² DRAM
With Manufacturability and Enhanced Cell Efficiency

Ki-Whan Song, Jin-Young Kim, Jae-Man Yoon, Sua Kim, Huijung Kim, Hyun-Woo Chung, Hyungi Kim, Kanguk Kim, Hwan-Wook Park, Hyun-Chul Kang, Nam-Kyun Tak, Dukja Park, Woo-Soon Kim, Member, IEEE, Yeong-Tae Lee, Yong-Chul Oh, Giyo-Young Jun, Jee-Soon Yoo, Dongjun Park, Senior Member, IEEE, Kyungseok Oh, Changhyun Kim, Senior Member, IEEE, and Young-Hyun Jun
Today’s Lecture: DRAM

- DRAM, Xilinx, and You
- DRAM: Bottom-up
- DRAM: Top-down
So, we amortize the edge circuits over big arrays.

People buy DRAM for the bits. “Edge” circuits are overhead.

“Word Line” “Row”
A “bank” of 128 Mb (512Mb chip -> 4 banks)

In reality, 16384 columns are divided into 64 smaller arrays.

13-bit row address input

8192 rows 16384 columns

134 217 728 usable bits (tester found good bits in bigger array)

16384 bits delivered by sense amps

Select requested bits, send off the chip

Recall DRAM Challenge #3b: Sensing

How do we reliably sense a 60mV signal?

Compare the bit line against the voltage on a “dummy” bit line.

“Dummy” bit line.
Cells hold no charge.

Bit line to sense
“sense amp”

Dummy bit line
"Sensing" is row read into sense amps

Slow! This 2.5ns period DRAM (400 MT/s) can do row reads at only 55 ns (18 MHz).

DRAM has high latency to first bit out. A fact of life.

13-bit row address input

1 of 8192 decoder

8192 rows 134 217 728 usable bits (tester found good bits in bigger array)

16384 columns

16384 bits delivered by sense amps

Select requested bits, send off the chip

An ill-timed refresh may add to latency

Parasitic currents leak away charge.

Solution: "Refresh", by rewriting cells at regular intervals (tens of milliseconds)

Diode leakage...
Latency is not the same as bandwidth!

Thus, push to faster DRAM interfaces.

What if we want all of the 16384 bits? In row access time (55 ns) we can do 22 transfers at 400 MT/s. 16-bit chip bus -> 22 x 16 = 352 bits << 16384. Now the row access time looks fast!

13-bit row address input

16384 columns

8192 rows 134 217 728 usable bits (tester found good bits in bigger array)

16384 bits delivered by sense amps

Select requested bits, send off the chip

Sadly, it’s rarely this good ...

What if we want all of the 16384 bits? The “we” for a CPU would be the program running on the CPU.

Recall Amdahl’s law: If 20% of the memory accesses need a new row access ... not good.

16384 columns

8192 rows 134 217 728 usable bits (tester found good bits in bigger array)

16384 bits delivered by sense amps

Select requested bits, send off the chip
**DRAM latency/bandwidth chip features**

- **Columns**: Design the right interface for CPUs to request the subset of a column of data it wishes:
  - 16384 bits delivered by sense amps
  - Select requested bits, send off the chip

- **Interleaving**: Design the right interface to the 4 memory banks on the chip, so several row requests run in parallel.
  - Bank 1
  - Bank 2
  - Bank 3
  - Bank 4

---

**Off-chip interface for the Micron part ...**

- A clocked bus: 200 MHz clock, data transfers on both edges (DDR).
- DRAM is controlled via commands (READ, WRITE, REFRESH, ...)
- Synchronous data output.

---

*Note! This example is best-case! To access a new row, a slow ACTIVE command must run before the READ.*
Opening a row before reading ...

Auto-Precharge READ

However, we can read columns quickly

Note: This is a “normal read” (not Auto-Precharge). Both READs are to the same bank, but different columns.
Why can we read columns quickly?

Column reads select from the 16384 bits here

16384 columns

8192 rows

134,217,728 usable bits
(tester found good bits in bigger array)

16384 bits delivered by sense amps

Select requested bits, send off the chip

Interleave: Access all 4 banks in parallel

Interleaving: Design the right interface to the 4 memory banks on the chip, so several row requests run in parallel.

Bank a  Bank b  Bank c  Bank d

Can also do other commands on banks concurrently.
DRAM controllers: reorder requests

(A) Without access scheduling (56 DRAM Cycles)

(B) With access scheduling (19 DRAM Cycles)

DRAM Operations:

- **P**: bank precharge (3 cycle occupancy)
- **A**: row activation (3 cycle occupancy)
- **C**: column access (1 cycle occupancy)

From: Memory Access Scheduling

Present and Future ...
MacBook Air ... too thin to use DIMMs

Mainboard: fills about 25% of the laptop

35 W-h battery: 63% of 2006 MacBook’s 55 W-h
Main board

Core i5: CPU + DRAM controller

4GB DRAM soldered to the main board

3-D memory stack

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM die size</td>
<td>10.7 mm × 13.3 mm</td>
</tr>
<tr>
<td>DRAM die thickness</td>
<td>50 µm</td>
</tr>
<tr>
<td>TSV count in DRAM</td>
<td>1,560</td>
</tr>
<tr>
<td>DRAM capacity</td>
<td>512 Mbit/die × 2 strata</td>
</tr>
<tr>
<td>CMOS logic die size</td>
<td>17.5 mm × 17.5 mm</td>
</tr>
<tr>
<td>CMOS logic die thickness</td>
<td>200 µm</td>
</tr>
<tr>
<td>CMOS logic bump count</td>
<td>3,497</td>
</tr>
<tr>
<td>CMOS logic process</td>
<td>0.18 µm CMOS</td>
</tr>
<tr>
<td>DRAM-logic FTI via pitch</td>
<td>50 µm</td>
</tr>
<tr>
<td>Package size</td>
<td>33 mm × 33 mm</td>
</tr>
<tr>
<td>BGA terminal</td>
<td>520 pin / 1 mm pitch</td>
</tr>
</tbody>
</table>

A 3D Stacked Memory Integrated on a Logic Device Using SMAFTI Technology

Yoshiki Kurita, Yuichiro Kikuchi, Shinya Morishita, Yoichiro Kurita, Masato Kikuchi, hideki Kikuchi, Koji Sakazawa, and Atsuhiko Kurita
Mitsubishi Electric Corporation, 1130 Motobu, Sagamihara, Kanagawa 229-1198, Japan

y.kurita@necel.com

Total Power=2W

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM</td>
<td>62.7°C</td>
</tr>
<tr>
<td>Logic</td>
<td>65.1°C</td>
</tr>
<tr>
<td>1m/s</td>
<td></td>
</tr>
<tr>
<td>65°C</td>
<td></td>
</tr>
<tr>
<td>70°C</td>
<td></td>
</tr>
<tr>
<td>40°C</td>
<td></td>
</tr>
<tr>
<td>25°C</td>
<td></td>
</tr>
</tbody>
</table>