Semiconductor Memories

Prof. MacDonald
Types of Memories

- **Volatile Memories**
  - require power supply to retain information
  - dynamic memories
    - use charge to store information and require refreshing
  - static memories
    - use feedback (latch) to store information – no refresh required

- **Non-Volatile Memories**
  - ROM (Mask)
  - EEPROM
  - FLASH – NAND or NOR
  - MRAM
Memory Hierarchy

- RF: 100pS, 100’s of bytes
- L1 SRAM: 1nS, 10’s of Kbytes
- L2 SRAM: 10nS, 100’s of Kbytes
- L3 DRAM: 100nS, 100’s of Mbytes
- Disks / Flash: 1us, Gbytes
Memory Hierarchy

- Large memories are slow
- Fast memories are small
- Memory hierarchy gives us illusion of large memory space with speed of small memory.
  - temporal locality
  - spatial locality
Register Files

- Fastest and most robust memory array
- Largest bit cell size
- Basically an array of large latches
- No sense amps – bits provide full rail data out
- Often multi-ported (i.e. 8 read ports, 2 write ports)
- Often used with ALUs in the CPU as source/destination
- Typically less than 10,000 bits
  - 32 32-bit fixed point registers
  - 32 60-bit floating point registers
SRAM

- Same process as logic so often combined on one die
- Smaller bit cell than register file – more dense but slower
- Uses sense amp to detect small bit cell output
- Fastest for reads and writes after register file
- Large per bit area costs
  - six transistors (single port), eight transistors (dual port)
- L1 and L2 Cache on CPU is always SRAM
- On-chip Buffers – (Ethernet buffer, LCD buffer)
- Typical sizes 16k by 32
Static Memory Cell

True Bit Line

T5

T2

T1

T3

T6

Complement Bit Line

Wordline
SRAM Interface

- Dirt simple
  - clk – only required if registered inputs and/or outputs
  - csn – usually negatively active chip select
  - wen – usually negatively active write enable
  - a – address bus – logarithmically related to number of locations
  - d – data in bus – inputs used during writes
  - q – data out bus – outputs used during reads

- read – csn active / wen inactive
  - data in the byte located by A bus launches out Q bus

- write – csn active / wen active
  - D bus value is loaded into memory location determined by A
SRAM Operation
Dual Port SRAM Operation

- Two complete single port interfaces
  - can read two locations simultaneously
  - can write two different locations simultaneously
  - can not write one location with two different ports
  - Typically requires 2 additional transistors per bit
    - (6 to 8 for 33% increase)
  - Typically requires 4 bit lines and 2 word lines
    - double the global routing than single port
Dual Port SRAM Interface

- Two merged SRAM interfaces
  - clk – if registered inputs and/or outputs
  - csnA / csnB – usually negatively active chip select
  - wenA / wenB – usually negatively active write enable
  - aA / aB – address bus – logarithmically related to number of locations
  - dA / dB – data in bus – inputs used during writes
  - qA / aB – data out bus – outputs used during reads
Two Port Memory Cell
Memory Compilers

- ASIC library providers give logic designers:
  - Standard cells - ANDs, NANDs, FLIP-FLOPs, etc
  - Chip IOs
  - Memory compilers
    - single port SRAM,
    - dual port SRAM, and
    - register files.

- Dial in size and generates verilog, layout, timing.

- DRAM, FLASH not supported
Dynamic RAM

- Most dense RAM (1 Gbit chips available)
- Historically, different semiconductor process so built on a separate die
- L3 Cache (old days) and computer main memory
- Requires refresh of data due to leakage
- New push to combine DRAM and logic
  - embedded DRAM, eDRAM
  - business case hard to close – yields drop
DRAM Bit Cells (1T)

DRAM used since the early 70s
Destructive Read
Highest density
DRAM Bit Cells (3T)

DRAM used in the early 70s
Non-destructive read

write bitline
write wordline
read bitline
read wordline
DRAM Cross Section
DRAM Interface Evolution

- Asynch DRAM – up until early 90’s
- Synch DRAM – add a clock
- Double Data Rate (DDR) SDRAM
- DDR2 SDRAM – Faster
- Others
  - RAMBUS – Intel supported – dead except for PS3
  - graphics versions – specifically for frame buffers
  - Mobile SDRAM – low power, good for palm pilots
DRAM Read Timings

RAS

CAS

Addr

Addr

RA1 CA1 RA CA

data 1 data 2
DRAM Read Timings (EDO)

RAS

CAS

Addr

RA1  CA1  CA

Addr

data 1  data 2
SDRAM

- Same array as Asynch DRAM
- Add pipelining to data to increase bandwidth
- Requires new clock signal
- Treats RAS, CAS, WE as command inputs
- Replaced by DDR for high performance apps
- Still alive for mobile applications due to power
## SDRAM Commands

<table>
<thead>
<tr>
<th>Command</th>
<th>CKEn.1</th>
<th>CKEn</th>
<th>CS</th>
<th>RAS</th>
<th>CAS</th>
<th>WE</th>
<th>DQM</th>
<th>ADDR</th>
<th>A10/</th>
<th>AP</th>
<th>BA</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode Register Set</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OP code</td>
</tr>
<tr>
<td>No Operation</td>
<td>H</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bank Active</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>H</td>
<td>X</td>
<td>RA</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>X</td>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read with Autoprecharge</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>X</td>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Write</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>X</td>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Write with Autoprecharge</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>X</td>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precharge All Banks</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>X</td>
<td>H</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precharge selected Bank</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>X</td>
<td>L</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burst Stop</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DQM</td>
<td>H</td>
<td></td>
<td></td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auto Refresh</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burst-Read-Single-WRITE</td>
<td>H</td>
<td>X</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>A9 Pin High (Other Pins OP code)</td>
</tr>
</tbody>
</table>

**Self Refresh**

- **Entry**: H L L L L H X
- **Exit**: L H H X X X X

**Precharge power down**

- **Entry**: H L H X X X X
- **Exit**: L H H H H X

**Clock Suspend**

- **Entry**: H L H X X X X
- **Exit**: L V V V V X
Maskable ROM

- Most dense ROM
- Tie bit high or low at mask level (metallization)
- Mistakes take weeks to fix with new silicon
  - hold lots at metal layer for quick implementation
**SDRAM Timings**

**Random READ Accesses**

*command* is just the value of CSN, CAS, RAS and WEN together in that cycle.

*address* is usually 9-11 bits plus 2 bits for bank address.
ROM/EEPROM/FLASH

- Metal Mask ROM
- Electrically erasable programmable ROM
- FLASH is block erasable only EEPROM
- EEPROM can be byte-written but requires extra transistor
- FLASH may take over the world – replacing disk drives with FLASH drives (no moving parts – more reliable).
- Motorola is leader in combining FLASH with logic
- Intel leader in NOR Flash
- Toshiba / Samsung leaders in NAND Flash
EPROM

- Erasable Programmable ROM
- Can be erased by UV light
- Programmed by Hot Carrier Injection
- Obsolete but still mentioned
  - only used in EE2369 to provide historical perspective
NOR ROM Structure
NOR ROM Structure
NAND ROM Structure
Flash Cross Section
FLASH

Semiconductor Memory Market

- DRAM
- FLASH
- SRAM

Forecast of Web Feet Inc.
FLASH

- **NOR Flash**
  - less dense (256 Mbit) but provides fast random read access
  - Erase FN / Program HEI
  - 100,000 write cycles
  - Slow erase, fast program and read
  - SRAM like interface – give an address – get a byte of data
  - great for code memory (bios, boot-up, cell phone, etc)

- **NAND Flash**
  - More than 2X denser – up to 2Gbit
  - Erase FN/ Program FN
  - Fast erase, slow program and read
  - 1,000,000 write cycles
  - IO like interface – not as simple as NOR
  - good for data storage – memory cards, IPODs, USB keydrives
### Flash Cross Section

<table>
<thead>
<tr>
<th></th>
<th>NAND</th>
<th>NOR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell Array</strong></td>
<td><img src="image1.png" alt="NAND Cell Array Diagram" /></td>
<td><img src="image2.png" alt="NOR Cell Array Diagram" /></td>
</tr>
<tr>
<td><strong>Layout</strong></td>
<td><img src="image3.png" alt="NAND Layout Diagram" /></td>
<td><img src="image4.png" alt="NOR Layout Diagram" /></td>
</tr>
<tr>
<td><strong>Cross-section</strong></td>
<td><img src="image5.png" alt="NAND Cross-section Diagram" /></td>
<td><img src="image6.png" alt="NOR Cross-section Diagram" /></td>
</tr>
<tr>
<td><strong>Cell size</strong></td>
<td>$4F^2$</td>
<td>$10F^2$</td>
</tr>
</tbody>
</table>
## NOR FLASH

**Table 10.4** Bias conditions of the NOR cells for erase, programming, and read operations

<table>
<thead>
<tr>
<th>Signal</th>
<th>Erase</th>
<th>Programming</th>
<th>Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bit line 1</td>
<td>Open</td>
<td>6 V</td>
<td>1 V</td>
</tr>
<tr>
<td>Bit line 2</td>
<td>Open</td>
<td>0 V</td>
<td>0 V</td>
</tr>
<tr>
<td>Source line</td>
<td>12 V</td>
<td>0 V</td>
<td>0 V</td>
</tr>
<tr>
<td>Word line 1</td>
<td>0 V</td>
<td>0 V</td>
<td>0 V</td>
</tr>
<tr>
<td>Word line 2</td>
<td>0 V</td>
<td>12 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Word line 3</td>
<td>0 V</td>
<td>0 V</td>
<td>0 V</td>
</tr>
</tbody>
</table>

![Diagram of NOR Flash circuit]
NAND Flash Reading

Table 10.5  Bias conditions of the NAND cells for erase, programming, and read operations

<table>
<thead>
<tr>
<th>Signal</th>
<th>Erase</th>
<th>Programming</th>
<th>Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bit line 1</td>
<td>Open</td>
<td>0 V</td>
<td>1 V</td>
</tr>
<tr>
<td>Bit line 2</td>
<td>Open</td>
<td>'0 V</td>
<td>1 V</td>
</tr>
<tr>
<td>Select line 1</td>
<td>Open</td>
<td>5 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Word line 1</td>
<td>0 V</td>
<td>10 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Word line 2</td>
<td>0 V</td>
<td>10 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Word line 3</td>
<td>0 V</td>
<td>10 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Word line 4</td>
<td>0 V</td>
<td>10 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Word line 5</td>
<td>0 V</td>
<td>20 V</td>
<td>0 V</td>
</tr>
<tr>
<td>Word line 6</td>
<td>0 V</td>
<td>10 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Word line 7</td>
<td>0 V</td>
<td>10 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Word line 8</td>
<td>0 V</td>
<td>10 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Select line 2</td>
<td>Open</td>
<td>0 V</td>
<td>5 V</td>
</tr>
<tr>
<td>Source line</td>
<td>Open</td>
<td>0 V</td>
<td>0 V</td>
</tr>
<tr>
<td>p-well 2</td>
<td>20 V</td>
<td>0 V</td>
<td>0 V</td>
</tr>
<tr>
<td>n-sub</td>
<td>20 V</td>
<td>0 V</td>
<td>0 V</td>
</tr>
</tbody>
</table>
Tunneling vs Injection

- FN: Fowler-Nordheim
- ONO: Oxide/Nitride/Oxide

Over the barrier: $V_g = 20V$ to $V_g = 10V$
Charge Pumps

- Flash and EEPROM architectures need unavailable higher voltage for programming (+10v)
- Charge pumps can pump a cap to get high voltage
- DC to DC (higher) converter - without inductors
- Need to consider Vmax across any gate oxide
- Generally cannot provide much power (I*V)
- Charge pumps used for a lot of other things like overdrive voltages and PLLs
Basic charge pump concept
Staged Diode Charge Pump

![Diagram of a Staged Diode Charge Pump](image)

**Figure 1.** Dickson charge pump circuit with a resistive load.
Dickson Charge Pump

\[ V_{\text{in}} \xrightarrow{M_{d1}} V_1 \xrightarrow{M_{d2}} V_2 \xrightarrow{M_{d3}} V_3 \xrightarrow{M_{d4}} V_4 \xrightarrow{M_{d15}} V_{\text{out}} \]

\[ C \xrightarrow{C} \xrightarrow{C} \xrightarrow{C} \xrightarrow{C} \xrightarrow{C} \]

\[ \phi \xrightarrow{\phi} \]

\[ C_{\text{out}} \]
Improved versions
Clock booster
4 Phase Charge Pump

\[ V_{\text{in}} \rightarrow M_0 \rightarrow M_1 \rightarrow M_2 \rightarrow C_{b1} \rightarrow M_3 \rightarrow M_4 \rightarrow C_{b2} \rightarrow V_{\text{out}} \]

\[ \text{Clk}_1 \rightarrow C_1 \rightarrow \text{Clk}_3 \rightarrow C_2 \rightarrow \text{Clk}_1 \rightarrow C_3 \rightarrow \text{Clk}_4 \rightarrow C_{b2} \rightarrow \text{Clk}_4 \rightarrow C_{b1} \rightarrow \text{Clk}_2 \rightarrow C_1 \rightarrow \text{Clk}_1 \]
CAMs

- SRAM structure that will do a parallel compare against the contents to provide a hit signal for each row (value)
- Used by caches to find if data is in cache
- Also used for translation look-aside buffers
- Also used in switches / routers to check destination address in ethernet against list of addresses
- Ternary CAMs allow for bit masking the compare
Encoders / Decoders

3 to 8 line decode

Encoder

Decoder

<table>
<thead>
<tr>
<th>input</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>001</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>010</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>011</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>101</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>110</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
### Read Only Memories - ROM

<table>
<thead>
<tr>
<th>Address</th>
<th>Data Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 2 3 4 5 6 7</td>
<td></td>
</tr>
<tr>
<td>000</td>
<td>1 0 1 0 1 0 0 0</td>
</tr>
<tr>
<td>001</td>
<td>0 1 0 0 0 0 0 1</td>
</tr>
<tr>
<td>010</td>
<td>0 0 1 0 0 1 0 0</td>
</tr>
<tr>
<td>011</td>
<td>0 0 1 1 0 0 0 0</td>
</tr>
<tr>
<td>100</td>
<td>0 0 1 0 0 0 1 1</td>
</tr>
<tr>
<td>101</td>
<td>1 1 1 0 0 1 1 1</td>
</tr>
<tr>
<td>110</td>
<td>1 1 0 0 0 0 1 0</td>
</tr>
<tr>
<td>111</td>
<td>1 1 1 1 0 0 0 1</td>
</tr>
</tbody>
</table>

2^n addresses addressed by n bit
m output data bits
Read Only Memories - ROM

- Address input
- n to $2^n$ decoder
- $2^n$ word lines
- Memory array
- m data output
ROM memory array

- Word line 2^{N-1}
- Word line 2^{N-2}
- Word line 2
- Word line 1
- Word line 0

Programmed “one”
Programmed “zero”

Weak pull down

Bit lines

Data m-1, data m-2, ..., data2, data1, data0
Random Access Memory (RAM)

6 Transistor SRAM cell

Address input

n+c

n
c

n to 2^n decoder

sram array of 6-t cells

peripheral circuits – column mux, sense amps, write circuitry

data in
m

data out
m

word lines
SRAM Organization

- Blocks with unity aspect ratio
- Rows
- Columns
- IO
Static Memory Cell
4D - Static Memory Cell
SRAM Read Cross-Section

DRAMs precharge to half Vdd so PFET enable required as well

Set Sense Amp

Isolation Circuit

Precharge Circuit

Wordline

TBL

CBL

Cell

Bit Line Precharge

Bit Line Isolation

TSA

CSA
SRAM Isolation & Pre-charge Circuits

- Sense Amp
- Bit Switch Circuit
- Bit Line Isolation
- Pre-charge Circuit
- Cells
  - Bit Line Pre-charge

Diagram shows the connection and isolation of bit lines and pre-charge circuits in a SRAM configuration.
SRAM Sense Amplifier Circuit

DRAMs precharge to half Vdd so PFET enable required as well
SRAM Internal Memory Waveforms

- Clock
- Word line
- Isolation
- Set Sense Amp
- Sense Amp Output
- Data
SRAM Write Head Circuit

Write Enable

Data

Bit Line True

Bit Line Complement
SRAM Decode Circuit
SRAM Cell with Center GND Contact

- Vdd
- PFET diffusion
- Ground
- NFET diffusion
- Word line (Polysilicon)
- Bit line contacts
SRAM Cell with Shared Vdd Contact

- Bit line contacts
- NFET diffusion
- PFET diffusion
- Word line (Polysilicon)
- Vdd
- Ground
Split Word Line SRAM Cell

- Word line (Polysilicon)
- NFET diffusion
- PFET diffusion
- Bit line contact
- Ground
- Vdd
- Bit line contact
Bit Cell Analysis – Read Disturb

starts at 0v but will jump up. If it jumps too high, can flip the bit. T6 is often not min L to keep the jump low.

precharged to 1.8v
Bit Cell Analysis – Read Disturb

If low, right data node ($V_{rd}$) cannot exceed the threshold of T2 or bit may flip.

$$(K_{n6} / 2) (V_{dd} - V_{rd} - V_{tn})^2 = (K_{n5} / 2) (2 (V_{dd} - V_{tn}) V_{rd} - V_{rd}^2)$$

$$K_{n6}/K_{n5} < (2(V_{dd} - 1.5 V_{tn}) V_{tn}) / (V_{dd} - 2*V_{tn})^2$$
Bit Cell Analysis - Write

- Must ensure that write head circuit can over power cell by the end of the write cycle.
- The side of the bit cell with a 0 dominates the write transaction as the pass transistor is an NFET.
- When the word line asserts the write head circuit drives a zero on one of the two sides.
- The bit data in the cell must be brought below the threshold of the cross-coupled inverter to flip the bit.
Radiation (particularly in space – but occasionally on Earth) causes the generation of charge in circuits.

SOI technology helps as it shields transistors from charge in the bulk silicon.

The bit cell node has a capacitance and introduced charge will change the voltage at the node.

If the voltage swing exceeds the threshold of the cross-coupled inverter, the bit will flip (i.e. soft error).

Q_{crit} is charge required to flip bit.

Data is bad, but the bit cell still works (thus soft error).
Bit Cell Analysis – Soft Error

T1

True Bit Line

T5

T2

T3

T4

T6

Wordline

Complement Bit Line

constant current source turned on for time t

Qcrit = I * t
Redundancy and Repair

- DRAM and SRAMs often have extra rows, columns, IOs and / or Blocks to replace those that are bad.
- Two dimensional redundancy (i.e. rows and columns) provides good coverage with minimized.
- Rule of thumb (which changes over time) is currently that 2 Mbit memories and above will benefit in terms of yield if redundancy is included.
Redundancy and Repair

![Graph showing percent yield vs. average number of failing cells per chip for ECC only, Redundancy only, and Redundancy and ECC.]
Built-In Self Test

- Many register files and embedded SRAMs in ASICs now come with wrapper logic that will test the memory.
- The BIST includes muxes that all the test logic access to the ports of the memory.
- A BIST state machine generates address and data and writes to the ram, followed by reads and compares of the data.
- Some BIST also implement redundancy if necessary – This is often referred to as Built-In Self Repair.