Recall Basic Memory Definition

K x N

<table>
<thead>
<tr>
<th>Address[log2(K)-1:0]</th>
<th>Data[N-1:0]</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>E</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

K locations, N bits per location
Address bus has log2(K) address lines, data bus has N data lines.

Semiconductor Memory Device Architecture

- n×m Device
  - n inputs called "address lines"
  - m outputs called "data lines"

<table>
<thead>
<tr>
<th>A1</th>
<th>A0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2×4 Decoder

4 x 5 Memory
(4 locations, 5 bits per location).

Additional Control Lines

<table>
<thead>
<tr>
<th>Address[log2(K)-1:0]</th>
<th>K x N</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>M</td>
</tr>
<tr>
<td>OE</td>
<td>E</td>
</tr>
<tr>
<td>W</td>
<td>M</td>
</tr>
</tbody>
</table>

Chip Select – must be asserted before Memory will respond to read or write operation. If negated, data bus is high impedance.

OE – Asserted for read operation, Memory will drive data lines.

W – Asserted for a write operation (Memory inputs data from data pins, processor writes to memory).
Can see use of CS, W and OE signals.

### Pentium Memory System

### Timing Characteristics

Memory access time is the time from a valid address being placed on the address bus to valid data appearing on the data bus.

Memory write time is the time from a valid address being placed on the address bus to the capturing of the data bus value by the memory.

Faster is better!!!!

### Memory Vocabulary

- **ROM** – Read Only Memory - a type of memory that cannot be written, can only be read. Contents determined a manufacture time.  
  - ROM is non-volatile – contents remain even when power is off.
- **PROM** – Programmable ROM – a type of memory whose contents can be programmed by the user  
  - OTP – One Time Programmable, a PROM is OTP if contents can be programmed only once.
- **EEPROM** – Electrically Eraseable PROM – contents be erased electrically by the user.  
  - Memory is not alterable under ‘normal’ operation.
Memory Vocabulary

- RAM – Random Access Memory – memory that can be both read and written during normal operation.
  - Contents are non-volatile, will be lost on power off.
- SRAM – static RAM – has the following characteristics:
  - Read, Write operations take equal amounts of time
  - Access to any 'random' location takes same amount of time.
  - Fastest access time of memory types.
  - Basic memory cell is a latch, takes 6 transistors per memory bit.

Static RAM Cell

Cypress 8K x 8 SRAM

Multiple Chip Enables – more flexibility
Data bus is bi-directional (8 I/O lines).
Note: a more precise name for this memory is Asynchronous SRAM.
It has no clock input, any change on address inputs accesses a new location.
Read Cycle

Cycle Time – how fast can I start another operation?

Access Time – how fast is data ready?

For SRAMs, Cycle Time = Access Time, this is a feature of SRAMs.

Write Cycle

Write Timings

Notice that Write Cycle = Read Cycle.

All volatile RAM types have this feature.
Other SRAM Types

- **SSRAM - Synchronous SRAM**
  - Has a clock input
  - Address, data lines latched on clock edge
  - Can perform burst cycles
- **What is a burst cycle?**
  - After first data value is output based upon address, data values in successive locations are output without needing to change address bus
  - Internal counter used for address value

---

**SSRAM Normal Cycle vs Burst Cycle**

- Address ready, 1st data ready in 2 clocks, successive locations ready in successive clocks. This burst cycle is a 2-1-1-1 cycle. Diagram shows burst suspended for 1 clock between 3rd & 4th clock. Normal 2-1-1-1 burst takes 5 clocks.

---

**Why Burst Cycles?**

- Burst Cycles more efficient at block memory transfers (memory location is successive locations)
  - Normal cycle would take 8 clocks to transfer data from 4 locations.
  - 2-1-1-1 burst takes only 5 clocks.
- Why do we want to transfer blocks of memory efficiently?
  - For cache fill operations.
  - So what is a cache?
A Cache System

What is a Cache?

- The “closer” a memory is to a CPU, the faster the data transfer between CPU and Memory.
  - Can’t get any closer than on the same chip as the CPU!
- Need LOTS of memory
- Don’t have enough room to put all memory on same chip as CPU
  - Put some memory on same die as CPU, will not be able to hold all needed data/programs but will hold most frequently used data/programs
  - Will need to swap out some data if we don’t find what we need in the cache (a miss!).

Cache operations

- Hit – memory contents we need is located in cache.
- Miss – memory contents we need is not in cache
  - Check next level of memory to see if it is in the cache.
  - When we find it, copy in an entire block of memory because if we a particular memory location, chances are that we will need its neighboring locations also very soon.
- Transfers between 1st level cache and CPU are random accesses, single word transfers.
- Transfers between all other memory caches are block oriented – transfer multiple words at a time!
x86 Family Burst Cycles

- 8088 - 8 bit data bus, no burst cycles. Read Cycle, Write cycle each took 4 clocks.
- 8086 – 16 bit data bus, no burst cycles. Read Cycle, Write cycle each took 4 clocks.
- 386 – 32 bit bus, no burst cycles, Read Cycle, Write cycle each took 2 clocks.
- 486 - 32 bit bus, Normal Read/Write cycles took 2 clocks. Supports a burst of 2-1-1-1 (5 clocks to read 16 bytes! 32-bit bus transfers 4 bytes per bus cycle).
- Pentium – 64 bit bus, Normal Read/Write cycle takes 2 clocks. Supports a burst of 2-1-1-1 (5 clocks to read 32 bytes).

Pipelined Bursts

- A 486 can only do successive burst cycles that look like: 2-1-1-1;2-1-1-1:2-1-1-1
  - After one burst is over, must start a complete new burst.
- A Pentium can do pipelined burst cycles 2-1-1-1;1*-1-1-1;1*-1-1-1
  The next burst cycle picks up where the previous burst cycle left off!!! No need to supply a new address for the successive burst cycles.

Memory Hierarchy

Cost/Bit
- The further away, the cheaper per bit.
  - Registers
  - Cache
  - Main Memory
  - Fixed Disk

Access/Speed
- The further away, the slower the access time, the higher the capacity (density).
  - Tape
  - Floppy
  - Zip
  - CD-ROM
  - CD-RWR

Capacity
Main Memory and DRAM

- For main memory, we need a memory whose primary characteristics are:
  - Dense (lots of bits!!!)
  - Cheap (cause we don’t want to spend too much for our lots of bits!)
  - Good at block transfers, can be rotten at random accesses.
- How do we achieve density? (lots of bits?)
  - Fewer transistors per bit!
- Dynamic RAM (DRAM) - one transistor + capacitor per bit!!!!!

---

DRAM Memory Cell

- Word line
- Cs
- Bit line
- Memory value stored on capacitor (a very small capacitor...)
- Data value appears on bit line

---

DRAM Characteristics

- Very dense (high capacity). Cheap per bit.
- Slow for Random Access (access to any location)
  - Cycle time $>>$ access time, Read cycle time = Write cycle time.
- Has special access modes to speed block transfers
  - Important since transfers to DRAM in modern computer system is always block-oriented for cache fills.
- Only has half the address pins that you would expect
  - 1M x 8 DRAM has only 10 address pins instead of 20
  - Reduces package size, can pack more DRAM chips per unit area. Address values multiplexed between row/column addresses
A 1M x 16 DRAM (Micron Tech.)

- A0-A9 - 10 address pins
- DQ1-DQ16 - 16 data pins
- RAS# - row address strobe — asserted when address pins contain row address.
- CASL#, CASH# - column address strobe — asserted when address pins contain row address. Both need to be asserted for 16 bit transfers, only one for either high or low byte transfers.

Address Muxing

Assume a 16-bit data wide bus (D0-D15), and 20 address lines A20-A1 (no A0 pin since 16-bit-wide data bus).

How are addresses split between Row, Column?

- Varies the fastest.

DRAM Controllers

- External logic called a DRAM controller needed to interface to DRAMs.
  - DRAM interfacing more complex than SRAM
- Provides muxing of Address lines
- Assertion of RAS, CAS lines
- Also keeps DRAM contents refreshed
  - Capacitors tend to leak. Memory contents needs to continually accessed in order to keep contents valid.
  - Special 'refresh' cycles are supported by DRAMs to support refreshing all of bits in a row with one cycle
  - DRAM controller responsible for running refresh cycles.
Timing Values

- Random access – \( T_{rc} = 104\text{ns}, \ T_{rac} = 50\text{ ns} \) (access time from row address strobe assertion)
- Block Mode
  - Access time to data on row: 50 ns
  - To column locations on same row: 25 ns
- Block transfer much more efficient than random access.
DRAM Chip Generations

- Each time a new DRAM Chip generation is released, capacity goes up by 4x
  - 16K x 1 (late 70’s), 64K x 1, 256K x 1, 1M x 1, 4M x 1, 16M x 1, 64M x 1, 256M x 1 (current), 1G x 1 (in testing), 4G x 1 (in labs)
- Why? because of muxed addresses, 1 address pin is actually two address bits. Two address bits give 4X more locations.

DRAM Modules: SIMMs and DIMMs

- Mount Memory Device Packages on Circuit Boards to Conserve Space
- 30-Pin SIMM – First – Single Byte Access
  - Used in “Pairs” Since in x86 1 Word=16 bits
- 72-Pin SIMM – Four Byte Access
  - Need Pairs for Pentium+ Since 64 bit Data Bus
- 168-pin DIMM – Eight Byte Access

SIMM – Single In-Line Memory Module
DIMM – Dual In-Line Memory Module

SIMMs/DIMMs invented to get denser packaging for DRAMs on motherboards.
Memory Device Specification

- "Dimension" of Storage Cell Array
  - 8 Mb – Refers to Eight Mega-bits (not Mega-Bytes!)
    - \(8 \text{ Mb (lower case 'b')}\)
    - \(= (8)(1024)(1024)\text{bits}\)
    - \(= (1)(1024)(1024)\text{Bytes}\)
    - \(= 1\text{MB} \text{ (upper case 'B')}}\)

Single 16Mb Device can be Arranged as:

- 16M \times 1 \text{ bits}
- 4M \times 4 \text{ nybbles}
- 2M \times 8 \text{ bytes}
- 1M \times 16 \text{ (words in the case of x86)}

- Many Devices are Typically “byte-wide” Devices
  - \(N \times 8\)

DRAM Variations

- SDRAM – Synchronous DRAM – has clock, supports burst transfers
- DDR-SDRAM - Double Data Rate SDRAM
  - Data transferred on EACH clock edge
  - Double the data rate of SDRAM
- RDRAM – Rambus DRAM
  - Will discuss this in more detail later
  - High speed signaling interface to support very fast data transfers
  - Very high latency (long access time to first location), but very fast at transferring successive locations – claim is better at block mode transfers than SDRAM/DDR-SDRAM.
Dual Port Memories

Support simultaneous access via two ports (left/right sides).
Support simultaneous read access to different locations or same locations.
Supports simultaneous write access to different locations.
Simultaneous write access to same location is not supported.
One obvious application is multiple-processor systems, but useful in many other situations.

Flash RAM

- SRAM-like interface, density of SRAM
- Non-volatile (retains contents when power is off)
- Read Cycle time same as SRAM (10’s of nanoseconds)
- Write Cycle >> Read Cycle
  - write times in microseconds
  - Can write individual locations or blocks of locations
- Applications include smart cards (credit cards, medical history cards, etc)
- Intel is market leader, has a Flash RAM cell that stores 2-bits per cell (can sense 4 different voltage levels from cell).