Memory Systems

COMP375 Computer Architecture and Organization
“Thanks for the memories”

Bob Hope

“A clear conscience is usually the sign of a bad memory”

Steven Wright
Goals

• Understand the different memory technologies
• Be able to determine the appropriate memory technology for the specific application
Simple Model of Memory

• Each object is stored in one unique location
• Reads and writes to objects are atomic

• The system maintains this model for the programmer even though the implementation may be different
Atomic Actions

• An atomic action is indivisible
• You can see the state of an object before or after an atomic action, but you cannot see any intermediate state
• If you atomically change a byte in memory, you will never catch it half way with only some of the bits changed
Single Assembler Statement

• If a memory change can be done in a single assembler statement, it can be atomic
  
  goat = 47;
  
  mov goat, 47

• If it takes multiple assembler instructions to do something, it will not be atomic
Key Aspects of Memory

• Technology
  – Properties of the underlying hardware

• Organization
  – The way the technology is used to form a working system.
  How to combine bits into a working system.
Characteristics of Memory

- Primary or secondary
- Random or sequential access
- Granularity of access
- Volatile or nonvolatile
- Read/write or read-only
Memory Hierarchy

- regs
- cache
- RAM
- DISK
- Removeable Media
Primary and Secondary Memory

• Primary memory holds the data and program so that it can be accessed by the processor. It is usually fast and volatile.
  – RAM
  – cache

• Secondary memory is typically used for long term storage. It is usually non-volatile and may be removable. Use I/O to access.
  – Hard Drive
  – CD
Which is the correct order of devices listed by speed?

1. internal disk drive
2. register
3. USB thumb drive
4. RAM

A. 1 2 3 4
B. 2 4 1 3
C. 2 1 4 3
D. 4 2 3 1
E. 4 2 1 3
Random Access

• Random access means that it takes the same length of time to read (or write) any data in the system
• The main computer memory is random access (RAM). The ROM is also random access.
• Some external memory devices, such as tapes, are strictly sequential access
• Some memory is a combination of random and sequential
Sequential Access
Addressing Granularity

• Some memory types (such as primary RAM) can be addressed by individual bytes

• Other types of memory, such as disks, can only be accessed in blocks
  – You cannot read or write less than a block
  – If you want to change one byte, you must read the block, change the byte and then write the block

• Rarely can you address individual bits
Volatility

• Most RAM in a computer will lose the data when the power is turned off
• Some memory systems, such as memory sticks or flash memory, can keep the data without power
• Some memory will maintain its data unless special systems are used
A hard drive is?

A. Volatile, random access, large granularity
B. Non-volatile, sequential access, byte granularity
C. Non-volatile, random access, large granularity
D. Non-volatile, sequential, large granularity
E. Non-volatile, random access, byte granularity
SRAM Technology

- Static RAM stores data in logic circuitry similar to a flip-flop
- It takes four to six transistors per bit
- High speed
- High power consumption and heat
- Frequently used for cache memory
SRAM Design

write enable

input

read

output
DRAM Technology

• Dynamic RAM stores the data as a charge in the capacitance of a single transistor
• Only one transistor is required per bit
• Slower than SRAM
• Heat and power consumption are less than SRAM
• Data values must be rewritten after reading. Write faster than read
DRAM Implementation

Address line

Transistor

Storage capacitor

Bit line B

Ground
DRAM Refreshing

• A capacitor gradually loses its charge.
• If left alone, the value of a bit in DRAM would be lost in less than a second.
• A refresh circuit periodically reads the data and writes it back to the DRAM.
Read-Only Memory

• A Read Only Memory chip (ROM) contains data that cannot be changed by an executing program
• Most PCs have a Basic Input/Output System (BIOS) program in ROM. This provides a set of functions to perform low level activities and initialization.
• Embedded systems often have all programs in some form of ROM
Programming a ROM

• The data in a ROM can be set by:
  – Manufactured to contain specific data
  – Destroying zero value bits by applying too much current to that bit to “blow the fuse”
  – Some programmable ROMs (PROM) can be changed by special machines that use higher voltages. Some PROMs can be erased by ultraviolet light
  – Electrically Erasable Programmable ROM (EEPROM) requires special circuitry to write which takes much longer than reading
In what type of memory would you put the program for an embedded device?

1. ROM
2. SRAM
3. EPROM
4. core
Flash Memory

- EEPROM used in thumb drives, cell phones, digital cameras and solid state disks (SSD)
- Random access read and first write
- Must be erased in blocks before rewriting
- All types of flash memory and EEPROM wear out after a certain number of erase operations
Memory Performance

• Latency
  – Time between the initiation of a request until the data is returned

• Transfer time
  – The time per byte once data starts moving

• Cycle Time
  – Measures how quickly the memory system can handle successive requests
  – The memory system may require additional time between memory accesses
Performance Comparison

- A disk drive has significant latency, but a high transfer rate
- A USB thumb drive has low latency, but a slow transfer rate
Synchronized Memory

• The memory system uses a clock and the CPU uses a different clock
• The difference in clocks may cause either the CPU or memory to pause briefly
• Synchronized clock systems align both clocks.
• Synchronized RAM available
  – SDRAM – Synchronized DRAM
  – SSRAM – Synchronized SRAM
Faster Memory Clocks

• Some memory systems run at twice the normal clock rate
• These memories are known as “Double Data Rate”
• DDR-DRAM – Double Data Rate DRAM
• DDR-SDRAM – Double Data Rate Synchronized DRAM
# Summary

<table>
<thead>
<tr>
<th>Type</th>
<th>Category</th>
<th>Erasure</th>
<th>Byte alterable</th>
<th>Volatile</th>
<th>Typical use</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRAM</td>
<td>Read/write</td>
<td>Electrical</td>
<td>Yes</td>
<td>Yes</td>
<td>Level 2 cache</td>
</tr>
<tr>
<td>DRAM</td>
<td>Read/write</td>
<td>Electrical</td>
<td>Yes</td>
<td>Yes</td>
<td>Main memory</td>
</tr>
<tr>
<td>ROM</td>
<td>Read-only</td>
<td>Not possible</td>
<td>NO</td>
<td>NO</td>
<td>Large volume appliances</td>
</tr>
<tr>
<td>PROM</td>
<td>Read-only</td>
<td>Not possible</td>
<td>NO</td>
<td>NO</td>
<td>Small volume equipment</td>
</tr>
<tr>
<td>EPROM</td>
<td>Read-only</td>
<td>UV light</td>
<td>NO</td>
<td>NO</td>
<td>Device prototyping</td>
</tr>
<tr>
<td>EEPROM</td>
<td>Read-mostly</td>
<td>Electrical</td>
<td>Yes</td>
<td>NO</td>
<td>BIOS</td>
</tr>
<tr>
<td>Flash</td>
<td>Read/write</td>
<td>Electrical</td>
<td>NO</td>
<td>NO</td>
<td>mobile device disks</td>
</tr>
</tbody>
</table>
Core Memory

Long ago, small iron rings were magnetized in either clockwise or counter-clockwise direction to store a bit.
Future Memory Systems

• Nonvolatile memory systems, such as flash memory, are becoming cheaper and bigger

• Disk drives are being replaced by nonvolatile integrated circuit memory

• Flash memory wears out faster than a hard drive

Flash memory is less bothered by vibrations
Consider a one bit memory cell. It has:

- input line for write
- output line for read
- R/W to indicate read or write
- select to activate this cell
Combining Bits

• Since the computer has more than one bit of memory, we need to organize the bits into memory units.
• The number of bits in a memory unit depends on how many bits will fit.
• Bits can be grouped into units such as bytes or words.
• If you have $X$ bits, you can divide it into $N$ groups of $M$ bits each where $X = N \times M$. 
Usual Organization

• DRAM chips are often created with small groups
• A 1Mbit DRAM chip might load or store a single bit at a time
• SRAM chips often input and output bytes or larger groups, such as words
A memory module connects to a 128 bit bus and has 8 chips. How many bits are stored in parallel on each chip?

A. 1  
B. 8  
C. 16  
D. 32  
E. 64
1 by N bit memory

Decoder

address

log₂

write

read

R/W

...
1 by N bit memory

Write data and control lines have been omitted for clarity.
2 by N bit memory
M by N bit memory

Decoder

address

$\log_2$
Memory Modules

What inputs and outputs does this memory module have if it contains 16K bytes?
How many address lines are required by a 16 KByte chip?

A. 8
B. 14
C. 16
D. 24
E. 16,384
16K Byte Module Pins

• Addressing 16K bytes requires 14 pins
• Data input and output could be shared on 8 pins
• Read/Write control is one pin
• Select is one pin
• The electronics may require additional pins for power, ground and clock
Multiple Memory Modules

- Assume we want to combine our 16K byte modules to form 16K of 32 bit words.
Module Selection

• If you have $X$ memory modules (where $X$ is a power of 2), the upper $\log_2 X$ bits of the address can be used to select which memory module to use.
64K of 32 bit Words
Interleaved Memory

• The previous example used the upper bits of the address to select the memory module
• Using the lower bits to select the memory module means that sequential addresses are in different modules
• This can be advantageous if you are accessing sequential addresses and want to avoid memory access cycle delays
64K of 32 bit Words Interleaved
Bus Width

- The width of the bus or the number of data wires in the bus, determines how many bits can be transferred from the RAM to the processor at one time.
- Many buses are 64 to 256 bits wide.
- The RAM needs to be designed to provide 64 to 256 bits at a time.