MODULE 7: Memory Systems

Lecture 7.1
Memory Systems

Prepared By:
- Scott F. Midkiff, PhD
- Luiz A. DaSilva, PhD
- Kendall E. Giles, PhD

Electrical and Computer Engineering
Virginia Tech

Lecture 7.1 Objectives

- Define random access memory
- Describe the main tradeoffs in static versus dynamic RAMs
- Define read-only memory and identify its applications
- Enumerate the different types of ROM available
- Describe how a ROM can be used to implement logical and arithmetic operations
- Enumerate the components of a memory hierarchy
- Define hit rate, miss rate, hit time, and miss penalty
- Describe the types of locality

Types of Memory

- Random Access Memory (RAM)
  - Can be written to and read from
  - Also called primary or main memory
  - Volatile: data is lost when power is turned off
- Read Only Memory (ROM)
  - Can be read from but not written to
  - Stores critical information needed to run the system, such as microcode and the program used to boot the system
  - Non-volatile: always retains data

RAM

- Semiconductor memory that can be read from and written to
- “Random” refers to the fact that values can be accessed in the same amount of time regardless of their actual physical location
- The more RAM you have:
  - The more programs you can run at the same time
  - The more complex data you’ll be able to deal with in running these programs
  - The better performance you will experience

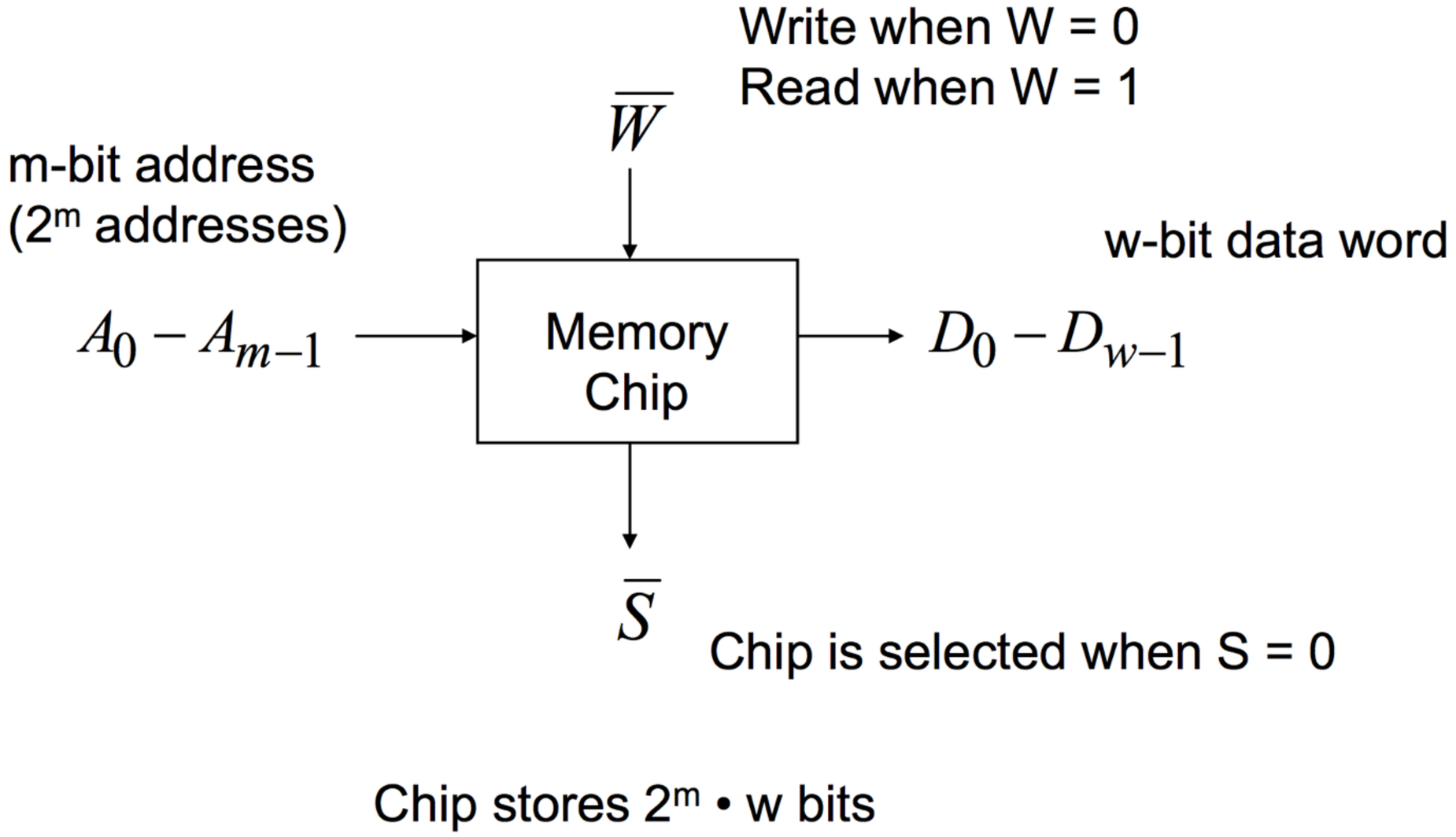
Static and Dynamic RAM (1)

- Static Memory (SRAM)
  - Based on flip-flops
  - Contents of each location persist as long as power is applied to the chip
- Dynamic Memory (DRAM)
  - Employ capacitors to store charge (charge level maps to a ‘0’ or a ‘1’)
  - Capacitors discharge with time, so DRAM must be refreshed frequently

Static and Dynamic RAM (2)

- SRAM
  - Faster
- DRAM
  - Can hold more information per unit area
  - Very economical, so widely employed (e.g., PCs)

Simplified Chip Pin-out



CHECK POINT

As a checkpoint of your understanding, please pause the video and make sure you can do the following:

- Define random access memory
- Describe the main tradeoffs in static versus dynamic RAMs

If you have any difficulties, please review the lecture video before continuing.

ROM

- Non-volatile memory that is pre-programmed with information
- ROM is used for:
  - Store microcode and low-level operating system routines
  - Implement combinational logic functions
  - Store programs in video games, calculators, automobile fuel injection controllers, etc.
- Several types of ROM
  - Factory-programmed
  - Programmable
  - Erasable

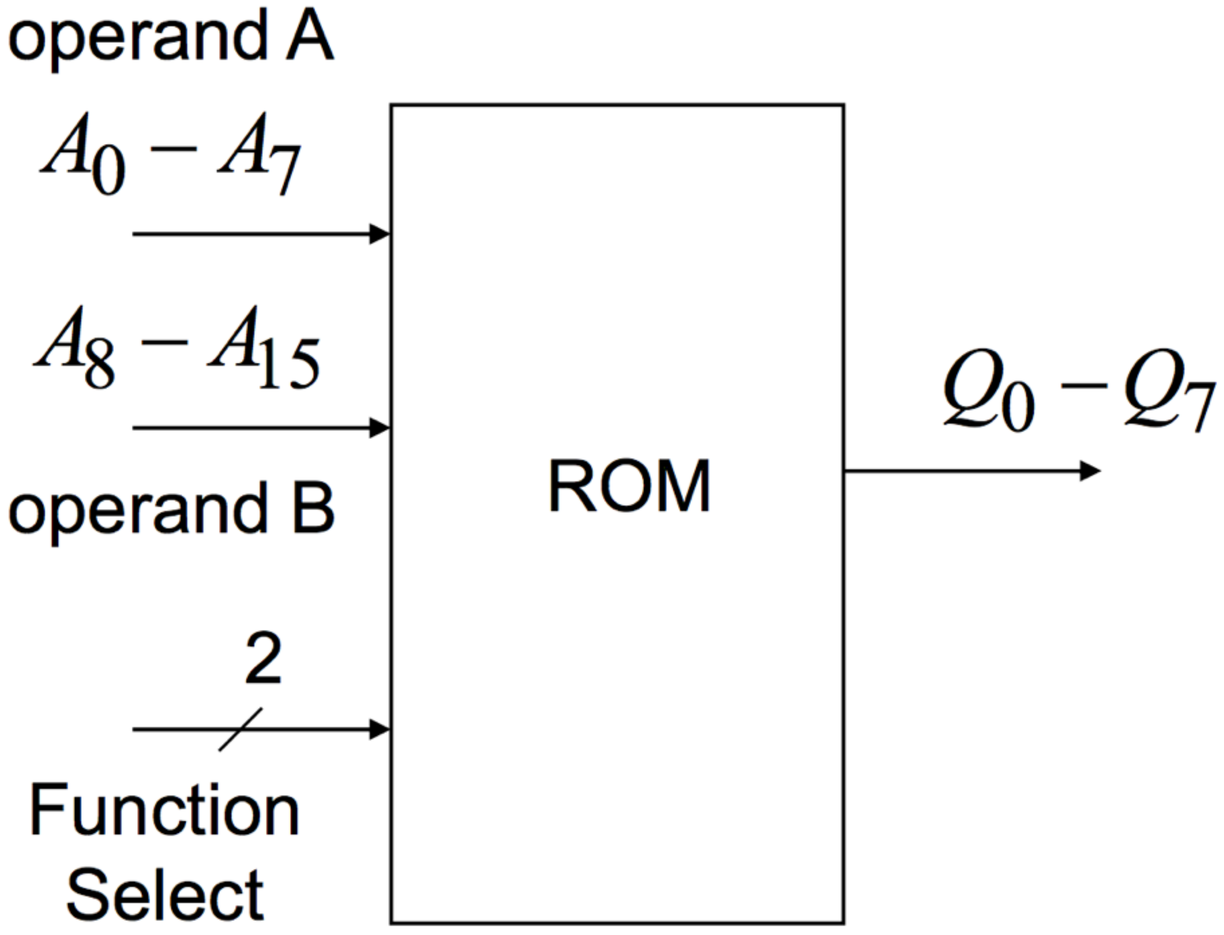
Types of ROM

- Mask-programmed ROMs – factory programmed
  - For high-volume applications
- Programmable ROMs – contents can be added w/ a PROM burner
  - For prototyping and for low-volume applications
- Erasable PROMs – can be repeatedly programmed and erased by the user
  - Flash EPROMs are very popular (information can only be erased by the block)
  - Electrically erasable PROMs (EEPROMs) are more costly, but information can be erased at the byte level

Example: Simple 8-bit ALU

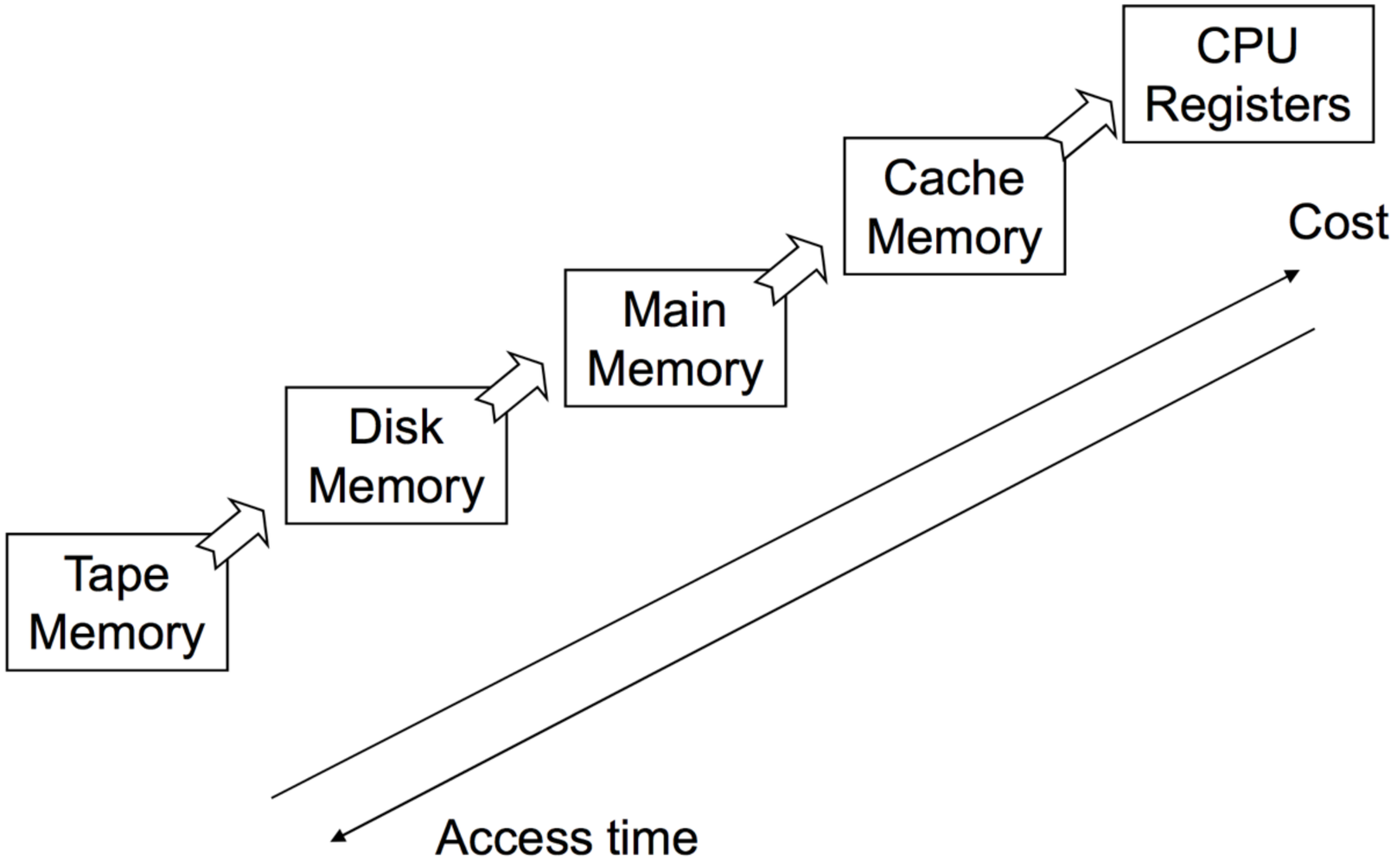
- Consider an Arithmetic Logic Unit (ALU) that performs 4 functions
  - Add, Subtract, Multiply, and Divide
- Implement using a look-up table
  - Table enumerates all possible combinations of operands and equivalent results from the 4 operations
  - Use operands & function select as addresses
  - Burn results into a ROM
- This approach is usually faster than implementing the functions in hardware (but doesn't scale very well)

Example: Look-up Table for ALU



A_{16}	A_{17}	Function
0	0	Add
0	1	Subtract
1	0	Multiply
1	1	Divide

Memory Hierarchy



CHECK POINT

As a checkpoint of your understanding, please pause the video and make sure you can do the following:

- Define read-only memory and identify its applications
- Enumerate the different types of ROM available
- Describe how a ROM can be used to implement logical and arithmetic operations
- Enumerate the components of a memory hierarchy

If you have any difficulties, please review the lecture video before continuing.

Primary, Secondary, Tertiary Memory

- Primary memory – RAM
  - Usually employ two levels of RAM (cache and main)
  - Small amount of faster, more expensive cache RAM
- Secondary memory – Disk drives
- Tertiary memory (off-line memory) – Tape, CDs/DVDs
- Instructions and data are continually passed from level to level during the course of processing

Hit and Miss

- A hit occurs when data is found at a given memory level
  - The hit rate is the percentage of time data is found at a given memory level
  - The hit time is the time required to access data at a given memory level
- A miss occurs when data is not found at a given memory level
  - The miss rate is the percentage of time data is not found at a given memory level
  - Miss rate = 1 - hit rate
  - The miss penalty is the time required to process a miss, including the time that it takes to replace a block of memory plus the time it takes to deliver the data to the processor

Hit and Miss Example

- To access a particular piece of data, the CPU first sends a request to its nearest memory, usually cache
- If the data is not in cache, then main memory is queried
- If the data is not in main memory, then the request goes to disk
- Once the data is located, then the data, and a number of its nearby data elements are fetched into cache memory

Locality

- An entire block of data is copied after a hit because the principle of locality tells us that once a byte is accessed, it is likely that a nearby data element will be needed soon
- Forms of locality:
  - Temporal locality: recently-accessed data elements tend to be accessed again
  - Spatial locality: accesses tend to cluster
  - Sequential locality: instructions tend to be accessed sequentially

CHECK POINT

As a checkpoint of your understanding, please pause the video and make sure you can do the following:

- Define hit rate, miss rate, hit time, and miss penalty
- Describe the types of locality

If you have any difficulties, please review the lecture video before continuing.

Summary

- RAM is semiconductor memory that can be read from and written to
  - Static (faster) and dynamic (more economical, needs to be refreshed periodically) are two types of RAM
- ROM is non-volatile memory that is pre-programmed with information
  - Flash memory is essentially EEPROM
- A computer implements a memory hierarchy, with CPU registers, cache, main memory, disk, and tape
  - A miss occurs when data is not found at a given memory level, and a block needs to be brought from the lower level of memory

MODULE 7: Memory Systems

Lecture 7.1
Memory Systems

Prepared By:
- Scott F. Midkiff, PhD
- Luiz A. DaSilva, PhD
- Kendall E. Giles, PhD

Electrical and Computer Engineering
Virginia Tech