Memories

Chapter 10

Physical Memory And Physical Addressing

Topics

- Introduction
- Static And Dynamic RAM
- Quantitative Measures Of Memory Technology
- Memory Controllers
- Synchronized Memory Technologies
- Multiple Data Rate Memory Technologies
- Examples Of Memory Technologies
- Memory Organization
- Memory Access And Memory Bus
- Memory Transfer Size

Topics

- Physical Addresses And Words
- Physical Memory Operations
- Word Size And Other Data Types
- An Extreme Case- Byte Addressing
- Byte Addressing With Word Transfers
- Using Powers Of Two
- Byte Alignment And Programming
- Memory Size And Address Space
- Programming With Word Addressing

Topics

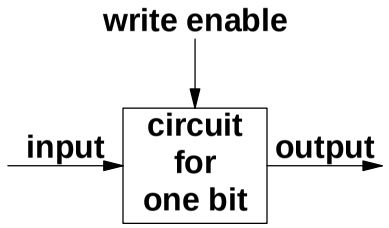
- Measures Of Memory Size
- Pointers And Data Structures
- A Memory Dump
- Memory Banks
- Interleaving
- Content Addressable Memory
- Ternary CAM
- Summary

Introduction

- This chapter discusses how basic memory system operates?
- Two main RAM technologies, SRAM and DRAM
- RAM
  * Random Access Memory
  * Primary memory system
  * Random access
  * Read-write capability
  * Volatile

Static RAM (SRAM)

- Stores each bit in a miniature digital circuits with multiple transistors



Miniature static RAM circuit that stores one data bit. The circuit contains multiple transistors.

- When write enable is on, circuits sets output = input
- When write enable is off, circuit ignores input, retains old output

Advantage And Disadvantages Of SRAM

- Advantage
  - high speed
- Disadvantage
  - power consumption and heat.
  - numerous transistors consume small amount of power each, but overall generate sizeable heat.

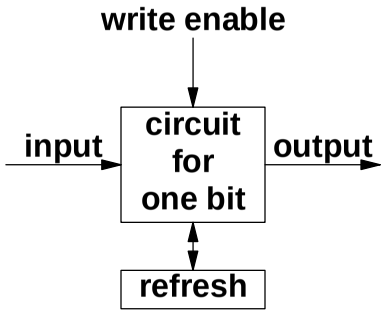
Dynamic RAM (DRAM)

- Uses a circuit that acts like a capacitor, with the device storing charge
- When value is written to DRAM, hardware charges or discharges according to the value.
- If a reading operation is performed, the charge is examined, and an equivalent digital value is generated

DRAM: Problem And Solution

- Problem
  - capacitor loses charge fast, sometimes in less than a second
- Solution
  - read a bit of memory before charge dissipates and write it back again
  - use a refresh circuit to perform the above task
  - a single refresh circuit cycles through one bit at a time
- Despite the refresh circuit DRAM is cheaper than SRAM

DRAM: Refresh Circuit



Bit in DRAM. An external refresh circuit must periodically read the data value and write it back again or the charge will dissipate, and the value will be lost.

Quantitative Measures Of Memory Technology

Density, Latency, Read and Write performance

- Density
  - the number of memory cells per square area of silicon
  - the number of bits that can be represented on a standard size chip.
  - high density: more memory in same space. Downside, more power generated and dissipated.
  - density increase follows Moore's law: doubles every 18 months.

Read And Write Performance

- Speed
  - how fast can memory respond to requests
  - cost of access (read) may be different from cost of update (write).
  - both read and write costs are important.

    In memory technologies, the time required to fetch information from memory differs from the time required to store information in memory, and the difference can be dramatic. Therefore, any measure of memory performance must give two values: the performance of read operations and the performance of write operations.

Latency And Memory Controllers

- Latency is the time that elapses between the start of an operation and the completion of an operation
- Memory Controller
  - found between the processor and physical memory
  - it translates memory address and requests
  - it responds to requests at the earliest, but may need more clocks to reset circuits

Because a memory system may need extra time between operations, latency is an insufficient measure of performance; a performance measure needs to measure the time required for successive operations

Read And Write Cycle Time

- To measure sequence of operations, compute average memory cycle time
  - Read cycle time (tRC)
  - Write cycle time (tWC)

  Read cycle time and write cycle time are used as measures of memory system performance because they measure how quickly the memory system can handle successive requests.

Synchronized Memory Technologies

- Clock controls when read or write operations begin
- What if the processor clock is different from memory clock?
  - controller can hold the request on either side for longer time
  - the difference in clock can affect performance
- Memory systems with synchronized clock system
  - SDRAM: Synchronized DRAM
  - SSRAM: Synchronized SRAM

Multiple Data Rate Memory Technologies

- Memory is usually the bottleneck

- Approach
  - lower memory clock times to some multiple of processor clock, say 2. Memory can deliver data faster.
- the underlying technology is called fast data range memories
- example of fast data rate memories are double data rate and quadruple data rate

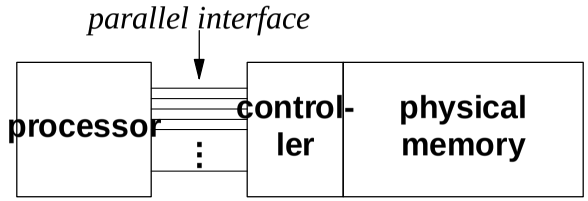
Examples Of Memory Technologies

Technology	Description
DDR-DRAM	Double Data Rate Dynamic RAM
DDR-SDRAM	Double Data Rate Synchronized Dynamic RAM
FCRAM	Fast Cycle RAM
FPM-DRAM	Fast Page Mode Dynamic RAM
QDR-DRAM	Quad Data Rate Dynamic RAM
QDR-SRAM	Quad Data Rate Static RAM
SDRAM	Synchronized Dynamic RAM
SSRAM	Synchronized Static RAM
ZBT-SRAM	Zero Bus Turnaround Static RAM
RDRAM	Rambus Dynamic RAM
RLDRAM	Reduced Latency Dynamic RAM

Examples of some of the commercially available RAM technologies

Memory Access And Memory Bus

- What is the connection structure between: processor-controller-memory
- Hardware connection between processor and memory is called a bus (memory bus)
- Bus provides parallel connections
- Many parallel wires between the processor and controller allow simultaneous transfers, i.e. higher performance



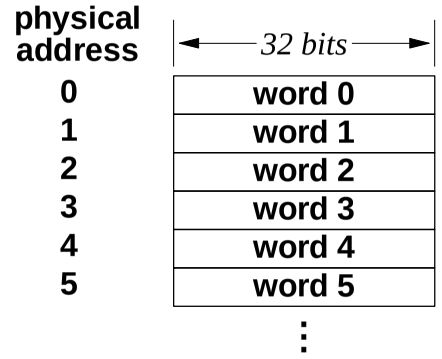
Parallel connections between a processor. A connection that contains N wires, allows N bits of data to be transferred simultaneously.

Memory Transfer Size

- Architecture view
  - Parallel connections improve performance
- Programming view
  - Parallel connections define a memory transfer size
  - Amount of data read or written in a single operation

Physical Addresses And Words

- Let N= memory transfer size, i.e. N bits per block (or word).
- Transfer size is also called the word size or width of word
- Each word is assigned an unique physical memory address
- The above approach is called word addressing
- Memory is an array of words



Physical memory addressing on a computer where word is 32 bits. We think of memory as an array of words.

Physical Memory Operations

- Controller supports read and write. In read, processor specifies address; in write, processor specifies address and data to be written.
- Controller accepts or delivers complete word only, no partial words.
- Physical memory is organized into words, where a word is equal to the memory transfer size. Each read or write operation applies to an entire word.

Word Size And Other Data Types

- Increasing the word size increases transfer rate, and hence higher performance
- Downside: a larger word increases cost of hardware
- What is an optimal word size?
  - large enough to hold integer and other common data values
  - should accommodate frequently used instructions
  - parallel hardware takes space and adds to cost, word size must be a compromise between performance and costs

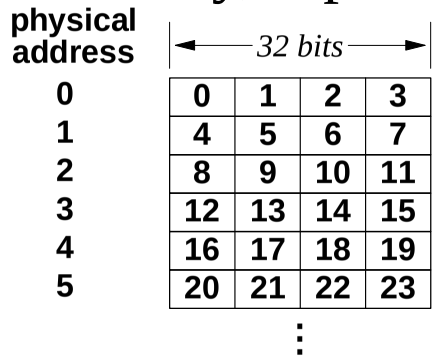
  32 bits is common now. Architects may choose memory size, integer single-precision FP and others to be same size i.e. 32 bits

An Extreme Case: Byte Addressing

- Programmers can access small data items like characters with byte addressing
- Memory is organized into array of words rather than bytes
- Consequences
  - more addresses are needed with byte addressing, since bytes are smaller than words
  - Controller must support byte transfer

Byte Addressing With Word Transfers

- Combining higher speed of word addressing, with convenience of byte addressing
  - Need intelligent memory controller that can translate between the two addresses
  - Controller accepts byte address from processor, but uses word address for memory, it performs a translation



A possible mapping between byte addresses used by a processor and word addresses used by the underlying hardware.

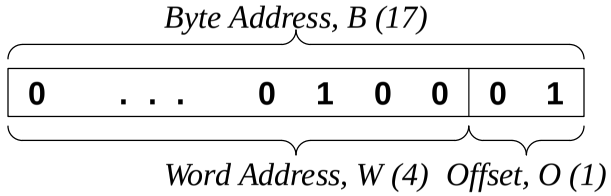
Byte Write Is Expensive

For writing a byte, controller reads entire word from memory, replaces the byte, and write the entire word back to memory

- Computation
  - Let B be byte address, W be word address, N be number of bytes/word
  - W = B / N (ignore remainder)
  - O = B mod N (byte offset)

Using Powers Of Two

- More computation means more ALU time and hardware
- Example: Let number of bytes/word, N = 4. Then,
  - W = all except last two bits in B
  - O = last two bits in B



Example of mapping from byte address 17 to word address 4 and offset 1. Using a power of two avoids arithmetic and logical calculations

To avoid arithmetic calculations such as division or reminder, physical memory is organized such that the number of bytes per word is a power of two, which means the translation from a byte address to word address and offset can be performed by extracting bits

Byte Alignment And Programming

- Byte alignment: integer value is aligned if the bytes of the integer correspond to a word of the underlying physical memory

  The organization of physical memory affects programming: even if a processor allows unaligned memory access, aligning data on boundaries that correspond to the physical word size can improve program performance

Memory Size And Address Space

- Processor limits address to be same size as integer
- If size of integer is 32, a 32 bit value represents 2^{32}
- You can have 2^{32} unique addresses
- Address space denotes set of possible addresses

Programming With Word Addressing

- If word address is used, manipulations will have to be performed for byte operations
- Logical bit shifts and bit masking are used if word/byte are powers of two
  - e.g. to extract leftmost byte (w >> 24) & 0xff

Measures Of Memory Size

Physical memory is organized into a set of M words that each contains N bytes; to make controller hardware efficient, M and N are each chosen to be powers of two.

- Exceptions
  - Kbs Kilo bits $= 2^{10}$ bits
  - Mbs Mega bits $= 2^{10}$ kbs

Pointers And Data Structures

- Memory addresses forms basis of abstractions like linked lists and trees.
- Pointer variable hold memory address
- Pointers can be assigned values and dereferenced

*int *iptr;*
*char *cptr;*
iptr is a pointer to a word
iptr++ increments value of iptr by 4 (assuming byte addressing).
iptr now moves to the next word
cptr++ increments value by 1, and cptr now points to next byte.

A Memory Dump

Address	Contents Of Memory			
0001bde0	00000000	0001bdf8	deadbeef	4420436f
0001bdf0	6d657200	0001be18	000000c0	0001be14
0001be00	00000064	00000000	00000000	00000002
0001be10	00000000	000000c8	0001be00	00000006

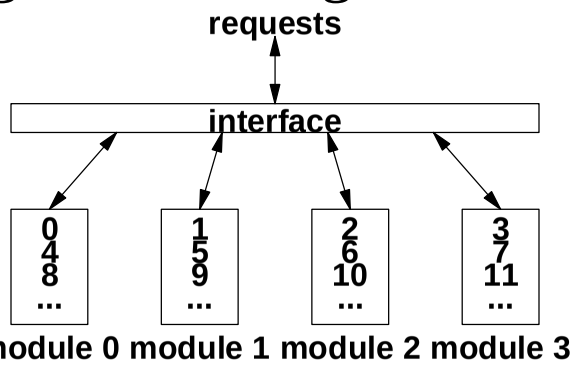
Small portion of a memory dump showing the contents of memory. The address column gives the memory address of the left-most byte on the line, and all values are shown in hexadecimal.

Memory Banks

- Processor uses multiple controllers to connect to multiple memory banks. Memory banks allow simultaneous parallel access
- Banks may be transparent to programmer, i.e. hardware automatically exploits parallelism
- Alternatively, programmer may be responsible for placing data on separate banks

Interleaving

- A memory optimization example
- Spreads consecutive memory bytes across modules
- Like banks, allows simultaneous access for obtaining higher performance
- Hardware deals with division of requests, the operations are hidden from programmers
- N-way interleaving means using N modules



4-way interleaving with numbers showing the bytes assigned to each module. Successive bytes are placed in seperate memory banks to optimize performance.

Content Addressable Memory

- An example of a memory that goes beyond storing data
- Combines technology and memory organization for high speed searching
- Conceptualize CAM as a memory organized in 2D array, each row is a slot.
- Processor specifi es the search key

Exact match

- Key matched against each slot
- Slots are same size, can be searched in parallel (number of slots does not affect performance)
- Parallel searches => more hardware
- CAM is expensive

Ternary CAM

- An alternate form of CAM
- Allows partial match searches on bits using 0 or 1 and not "don’t care"

Summary

- Examined technology and organization aspects of memory
- Physical memory are organized into words and accessed via controllers
- Programmers fi nd byte addressing convenient
- Byte addressing can be translated into word addressing and vice-versa
- Using word to be a multiple of two of byte size allows translation to
  be performed without computation (using shift and mask)
- Pointers allow programmers to obtain and manipulate memory addresses
- CAM combines memory technology and organization to provide high speed search mechanism