Chapter 7: Memory System Design

Topics

7.1 Introduction: The Components of the Memory System
7.2 RAM Structure: The Logic Designer's Perspective
7.3 Memory Boards and Modules

Fig 7.1 The CPU—Memory Interface

Sequence of events:
Read:
1. CPU loads MAR, issues Read, and REQUEST
2. Main memory transmits words to MDR
3. Main memory asserts COMPLETE

Write:
1. CPU loads MAR and MDR, asserts Write, and REQUEST
2. Value in MDR is written into address in MAR
3. Main memory asserts COMPLETE
Tbl 7.1  Some Memory Properties

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Intel 8088</th>
<th>Intel 8086</th>
<th>PowerPC 601</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>CPU word size</td>
<td>16 bits</td>
<td>16 bits</td>
<td>64 bits</td>
</tr>
<tr>
<td>m</td>
<td>Bits in a logical memory address</td>
<td>20 bits</td>
<td>20 bits</td>
<td>32 bits</td>
</tr>
<tr>
<td>s</td>
<td>Bits in smallest addressable unit</td>
<td>8 bits</td>
<td>8 bits</td>
<td>8 bits</td>
</tr>
<tr>
<td>b</td>
<td>Data bus size</td>
<td>8 bits</td>
<td>16 bits</td>
<td>64 bits</td>
</tr>
<tr>
<td>$2^m$</td>
<td>Memory word capacity, s-sized wds</td>
<td>$2^{20}$ words</td>
<td>$2^{20}$ words</td>
<td>$2^{32}$ words</td>
</tr>
<tr>
<td>$2^n x$s</td>
<td>Memory bit capacity</td>
<td>$2^{20} x 8$ bits</td>
<td>$2^{20} x 8$ bits</td>
<td>$2^{32} x 8$ bits</td>
</tr>
</tbody>
</table>

(Tab. 7.2 Memory Performance Parameters)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Units</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_a$</td>
<td>Access time</td>
<td>time</td>
<td>Time to access a memory word</td>
</tr>
<tr>
<td>$t_c$</td>
<td>Cycle time</td>
<td>time</td>
<td>Time from start of access to start of next access</td>
</tr>
<tr>
<td>k</td>
<td>Block size</td>
<td>words</td>
<td>Number of words per block</td>
</tr>
<tr>
<td>$\omega$</td>
<td>Bandwidth</td>
<td>words/time</td>
<td>Word transmission rate</td>
</tr>
<tr>
<td>$t_l$</td>
<td>Latency</td>
<td>time</td>
<td>Time to access first word of a sequence of words</td>
</tr>
<tr>
<td>$t_{bl}$</td>
<td>Block time</td>
<td>time</td>
<td>Time to access an entire block of words</td>
</tr>
<tr>
<td>$t_l + k/\omega$</td>
<td>access time</td>
<td></td>
<td>(Information is often stored and moved in blocks at the cache and disk level.)</td>
</tr>
</tbody>
</table>
## Table 7.3 The Memory Hierarchy, Cost, and Performance

<table>
<thead>
<tr>
<th>Component</th>
<th>CPU</th>
<th>Cache</th>
<th>Main Memory</th>
<th>Disk Memory</th>
<th>Tape Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>Random</td>
<td>Random</td>
<td>Random</td>
<td>Direct</td>
<td>Sequential</td>
</tr>
<tr>
<td>Capacity, bytes</td>
<td>64-1024+</td>
<td>8KB-8MB</td>
<td>64MB-2GB</td>
<td>8GB</td>
<td>1TB</td>
</tr>
<tr>
<td>Latency</td>
<td>.4-10ns</td>
<td>.4-20ns</td>
<td>10-50ns</td>
<td>10ms</td>
<td>10ms-10s</td>
</tr>
<tr>
<td>Block size</td>
<td>1 word</td>
<td>16 words</td>
<td>16 words</td>
<td>4KB</td>
<td>4KB</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>System clock rate</td>
<td>System Clock rate-80MB/s</td>
<td>50MB/s</td>
<td>1MB/s</td>
<td></td>
</tr>
<tr>
<td>Cost/MB</td>
<td>High</td>
<td>$10</td>
<td>$.25</td>
<td>$0.002</td>
<td>$0.01</td>
</tr>
</tbody>
</table>

*Some Typical Values:*†

†As of 2003-4. They go out of date immediately.

---

**Fig 7.3 Conceptual Structure of a Memory Cell**

Regardless of the technology, all RAM memory cells must provide these four functions: Select, DataIn, DataOut, and R/W.

This “static” RAM cell is unrealistic in practice, but it is functionally correct. We will discuss more practical designs later.
Fig 7.4 An 8-Bit Register as a 1-D RAM Array

The entire register is selected with one select line, and uses one R/W line.

Data bus is bidirectional and buffered. (Why?)

Fig 7.5 A 4 x 8 2-D Memory Cell Array

2-4 line decoder selects one of the four 8-bit arrays

R/W is common to all

Bidirectional 8-bit buffered data bus
Fig 7.6  A 64 K x 1 Static RAM Chip

~square array fits IC design paradigm

Selecting rows separately from columns means only 256 x 2 = 512 circuit elements instead of 65536 circuit elements!

CS, Chip Select, allows chips in arrays to be selected individually

This chip requires 21 pins including power and ground, and so will fit in a 22-pin package.

Fig 7.7  A 16 K x 4 SRAM Chip

There is little difference between this chip and the previous one, except that there are 4 64-1 multiplexers instead of 1 256-1 multiplexer.

This chip requires 24 pins including power and ground, and so will require a 24-pin package. Package size and pin count can dominate chip cost.
Fig 7.8 Matrix and Tree Decoders

- 2-level decoders are limited in size because of gate fan-in. Most technologies limit fan-in to ~8.
- When decoders must be built with fan-in >8, then additional levels of gates are required.
- Tree and matrix decoders are two ways to design decoders with large fan-in:

3-to-8 line tree decoder constructed from 2-input gates.

4-to-16 line matrix decoder constructed from 2-input gates.

Fig 7.9 Six-Transistor Static RAM Cell

This is a more practical design than the 8-gate design shown earlier.

A value is read by precharging the bit lines to a value 1/2 way between a 0 and a 1, while asserting the word line. This allows the latch to drive the bit lines to the value stored in the latch.
Access time from Address—the time required of the RAM array to decode the address and provide value to the data bus.

Write time—the time the data must be held valid in order to decode address and store value in memory cells.
**Fig 7.12 Dynamic RAM Cell Organization**

Capacitor will discharge in 4–15 ms.

Refresh capacitor by reading (sensing) value on bit line, amplifying it, and placing it back on bit line where it recharges capacitor.

**Write:** place value on bit line and assert word line.

**Read:** precharge bit line, assert word line, sense value on bit line with sense/amp.

This need to refresh the storage cells of dynamic RAM chips complicates DRAM system design.

**Fig 7.13 Dynamic RAM Chip Organization**

- Addresses are time-multiplexed on address bus using RAS and CAS as strobes of rows and columns.
- CAS is normally used as the CS function.

Notice pin counts:
- Without address multiplexing: 27 pins including power and ground.
- With address multiplexing: 17 pins including power and ground.
DRAM Read and Write Cycles

Notice that it is the bit line precharge operation that causes the difference between access time and cycle time.

DRAM Refresh and Row Access

- **Refresh** is usually accomplished by a “RAS-only” cycle. The row address is placed on the address lines and RAS asserted. This refreshes the entire row. CAS is not asserted. The absence of a CAS phase signals the chip that a row refresh is requested, and thus no data is placed on the external data lines.

- Many chips use “CAS before RAS” to signal a refresh. The chip has an internal counter, and whenever CAS is asserted before RAS, it is a signal to refresh the row pointed to by the counter, and to increment the counter.

- Most DRAM vendors also supply one-chip DRAM controllers that encapsulate the refresh and other functions.

- **Page mode**, **nibble mode**, and **static column mode** allow rapid access to the entire row that has been read into the column latches.

- **Video RAMS**, **VRAMS**, clock an entire row into a shift register where it can be rapidly read out, bit by bit, for display.
Types of High-Speed RAM

- FPM DRAM (Fast Page Mode DRAM)
- EDO DRAM (Extended Data Out DRAM)
- SDRAM (Synchronous DRAM)
- RDRAM (Rambus DRAM)
- DDR SDRAM (Double Data Rate SDRAM)
- SynchLink DRAM
- Dual Port Graphics Buffer
- SGRAM (Synchronous Graphics RAM)
- DDR SGRAM (Double Data Rate SGRAM)
- SSRAM (Synchronous SRAM)
- DDR SSRAM (Double Data Rate SSRAM)

FPM DRAM (Fast Page Mode DRAM)

- A DRAM provided with the page mode of higher-speed than that of the conventional DRAM.
- Although FPM DRAM executes data IO only once during one cycle of RAS# in random access, it can continuously execute data IO during one cycle of RAS# in the page mode.
- In the page mode, the access time of the second data and thereafter becomes faster.
- Other types of DRAM such as NB (Nibble Mode) and SC (Static Column Mode) that realize higher-speed using specifications different from those of FPM DRAM also exist.
- In 1995, however, FPM DRAM represented approximately 90% of all DRAM shipments.
**EDO DRAM (Extended Data Out DRAM)**

- Faster than FPM DRAM.
- If read cycle of FPM DRAM is made higher speed, the data output time becomes shorter.
- EDO DRAM provides with extended output functions, the data output time does not become short even in higher speed.
  - holds the data valid even after the signal which "strobes" the column address goes inactive.
  - allows faster CPU's to manage time more efficiently; i.e., while the EDO DRAM is retrieving an instruction for the microprocessor, the CPU can perform other tasks without concern that the data will become invalid.
- In addition, since EDO DRAM and FPM DRAM are compatible DRAM that have packages with the same pin configuration, EDO DRAM can easily replace FPM DRAM.
- In the middle of 1996, EDO DRAM represented approximately 50% of all the DRAM shipment.

**SDRAM (Synchronous DRAM)**

- A type of DRAM which operates in synchronization with input clock.
  - It latches each control signal at the rising edge of basic input clock and inputs/outputs data in synchronization with the clock signal.
- With synchronous control, the DRAM latches information from the processor under control of the system clock. These latches store the addresses, data and control signals, which allows the processor to handle other tasks. After a specific number of clock cycles the data becomes available and the processor can read it from the output lines.
- SDRAM represented more than 50% of DRAM shipments around 1998.
**DDR (Double Data Rate)**

- A technology designed to double the clock speed of the memory
- Activates output on both the rising and falling edge of the system clock rather than on just the rising edge, potentially doubling output.

- DDR SDRAM (Double Data Rate SDRAM)
- DDR SSRAM (Double Data Rate SSRAM)
- DDR SGRAM (Double Data Rate SGRAM)
Comparison - Access time of high-speed DRAM

- **Random access time**: Access time in which both the row address and the column address different from those of the preceding cycle are accessed.
- **Burst access time**: Access time in which the same row address and a column address different from that of the preceding cycle are accessed.

<table>
<thead>
<tr>
<th></th>
<th>Random Access Time (ns)</th>
<th>Burst Access Time (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPM DRAM</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>EDO DRAM</td>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td>SCRAM</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td>RDRAM</td>
<td>40</td>
<td>1.7 (1.3)</td>
</tr>
<tr>
<td>Conventional SRAM</td>
<td>15</td>
<td>-</td>
</tr>
<tr>
<td>Synchronous-type SRAM</td>
<td>6</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note: Figures in the parentheses show the value of Direct RDRAM.*

Memory Bus Clock Each High-speed DRAM Can Support
Fig 7.16  A 2-D CMOS ROM Chip

Tbl 7.4  ROM Types

<table>
<thead>
<tr>
<th>ROM Type</th>
<th>Cost</th>
<th>Programmability</th>
<th>Time to Program</th>
<th>Time to Erase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask-programmed ROM</td>
<td>Very inexpensive</td>
<td>At factory only</td>
<td>Weeks</td>
<td>N/A</td>
</tr>
<tr>
<td>PROM</td>
<td>Inexpensive</td>
<td>Once, by end user</td>
<td>Seconds</td>
<td>N/A</td>
</tr>
<tr>
<td>EPROM</td>
<td>Moderate</td>
<td>Many times</td>
<td>Seconds</td>
<td>20 minutes</td>
</tr>
<tr>
<td>Flash EPROM</td>
<td>Expensive</td>
<td>Many times</td>
<td>100 μs</td>
<td>1 s, large block</td>
</tr>
<tr>
<td>EEPROM</td>
<td>Very expensive</td>
<td>Many times</td>
<td>100 μs</td>
<td>10 ms, byte</td>
</tr>
</tbody>
</table>
Memory Boards and Modules

- There is a need for memories that are larger and wider than a single chip.
- Chips can be organized into “boards.”
  - Boards may not be actual, physical boards, but may consist of structured chip arrays present on the motherboard.
- A board or collection of boards make up a memory module.

Memory modules:
- Satisfy the processor–main memory interface requirements.
- May have DRAM refresh capability.
- May expand the total main memory capacity.
- May be interleaved to provide faster access to blocks of words.

Fig 7.17 General Structure of a Memory Chip
This is a slightly different view of the memory chip than previous.

Multiple chip selects ease the assembly of chips into chip arrays. Usually provided by an external AND gate.
**Fig 7.18 Word Assembly from Narrow Chips**

All chips have common CS, R/W, and Address lines.

- P chips expand word size from s bits to \( p \times s \) bits.

**Fig 7.19 Increasing the Number of Words by a Factor of \( 2^k \)**

The additional k address bits are used to select one of \( 2^k \) chips, each one of which has \( 2^m \) words:

- Word size remains at s bits.
Fig 7.20 Chip Matrix Using Two Chip Selects

This scheme simplifies the decoding from use of a (q+k)-bit decoder to using one q-bit and one k-bit decoder.

Multiple chip select lines are used to replace the last level of gates in this matrix decoder scheme.

Fig 7.21 Three-Dimensional Dynamic RAM Array

- CAS is used to enable top decoder in decoder tree.
- Use one 2-D array for each bit. Each 2-D array on separate board.
Fig 7.22  A Memory Module and Its Interface

Must provide—
- Read and Write signals.
- Ready: memory is ready to accept commands.
- Address—to be sent with Read/Write command.
- Data—sent with Write or available upon Read when Ready is asserted.
- Module select—needed when there is more than one module.

Bus Interface:

Control signal generator:
- for SRAM, just strobes data on Read, Provides Ready on Read/Write
- For DRAM—also provides CAS, RAS, R/W, multiplexes address, generates refresh signals, and provides Ready.

Fig 7.23  Dynamic RAM Module with Refresh Control

- Address
- Refresh clock and control
- Memory timing generator
- Module select
- Read
- Write
- Ready
- Data
- Refresh
- Grant
- Refresh counter
- Address register
- Address multiplier
- Board and chip selects
- RAS
- CAS
- R/W
- Dynamic RAM array
- Data lines
- Data register
Fig 7.24 Two Kinds of Memory Module Organiz’n.

Memory modules are used to allow access to more than one word simultaneously.

(a) Consecutive words in consecutive modules (interleaving)

(b) Consecutive words in the same module

---

Fig 7.25 Timing of Multiple Modules on a Bus

If time to transmit information over bus, $t_b$, is < module cycle time, $t_c$, it is possible to time multiplex information transmission to several modules;
Example: store one word of each cache line in a separate module.

**Main Memory Address:**

| Word | Module No. |

This provides successive words in successive modules.

**Timing:**

<table>
<thead>
<tr>
<th>Module 0</th>
<th>Write Module 3</th>
<th>……</th>
<th>Module 0 Data return</th>
</tr>
</thead>
</table>

With interleaving of $2^k$ modules, and $t_b < t_c/2^k$, it is possible to get a $2^k$-fold increase in memory bandwidth, provided memory requests are pipelined. DMA satisfies this requirement.
Memory System Performance

Breaking the memory access process into steps:

For all accesses:
- transmission of address to memory
- transmission of control information to memory (R/W, Request, etc.)
- decoding of address by memory

For a Read:
- return of data from memory
- transmission of completion signal

For a Write:
- transmission of data to memory (usually simultaneous with address)
- storage of data into memory cells
- transmission of completion signal

The next slide shows the access process in more detail.

Fig 7.26  Sequence of Steps in Accessing Memory

“Hidden refresh” cycle. A normal cycle would exclude the pending refresh step. -more-
Example SRAM Timings

Approximate values for static RAM Read timing:

- Address bus drivers turn-on time: 40 ns.
- Bus propagation and bus skew: 10 ns.
- Board select decode time: 20 ns.
- Time to propagate select to another board: 30 ns.
- Chip select: 20 ns.

PROPAGATION TIME FOR ADDRESS AND COMMAND TO REACH CHIP: 120 ns.

- On-chip memory read access time: 80 ns.
- Delay from chip to memory board data bus: 30 ns.
- Bus driver and propagation delay (as before): 50 ns.

TOTAL MEMORY READ ACCESS TIME: 280 ns.

Moral: 70 ns chips do not necessarily provide 70 ns access time!
Chapter 7 Summary

- Most memory systems are multileveled—cache, main memory, and disk.
- Static and dynamic RAM are fastest components, and their speed has the strongest effect on system performance.
- Chips are organized into boards and modules.
- Larger, slower memory is attached to faster memory in a hierarchical structure.