Basic Concepts

- Programs and data must be in main memory during execution
- Speed of operations is highly dependent upon the speed of transfer between the CPU and main memory
- Ideally main memory would be fast, large, and inexpensive (can not have all three)
- Max size of memory is determined by number of address bits, e.g. 16 bit addresses, \(2^{16} = 64K\) memory locations, 32 bit addresses \(2^{32} = 4G\) memory locations

Connecting Main Memory to the CPU

```
<table>
<thead>
<tr>
<th>CPU</th>
<th>Main Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR</td>
<td>Up to (2^k) addressable locations</td>
</tr>
<tr>
<td>MDR</td>
<td>Word length = (n) bits</td>
</tr>
<tr>
<td>k bit address bus</td>
<td>Control Lines (read, Write, MFC, …)</td>
</tr>
<tr>
<td>k bit data bus</td>
<td></td>
</tr>
</tbody>
</table>
```
Basic Concepts 2

• Memory Access Time
  – time that elapses between the initiation of an operation and the completion
  – e.g. the time between the Read and MFC signals

• Memory Cycle Time
  – minimum time delay required between the initiation of two successive memory operations
  – e.g. the time between two successive read operations

• Random-Access-Memory (RAM)
  – a memory unit in which any location can be accessed for a read or write operation in some fixed amount of time that is independent of the location

• Cache Memory
  – small, fast memory inserted between the CPU and main memory used to reduce memory access time
  – stores currently active program segments and data

• Memory Interleaving
  – organizing memory into a number of modules and arranges addressing so that successive words are placed in different modules
  – used when parallel access to the different modules is possible
Basic Concepts 3

• Virtual Memory
  – used to increase the apparent size of main memory
  – virtual address space is larger than physical memory size
  – disk storage is used to hold program segments and data which are not being currently used by the CPU
  – requires a special control circuit, a memory mapping unit
    • to translate virtual addresses used by the CPU to actual memory addresses
    • to determine if the necessary program segment or data is in main memory or in disk storage
  – significant time delay associated with bringing in data from disk storage
    • there are many techniques for minimizing the number of time this retrieval must occur
Semiconductor RAM Memories

- Introduced in the late 1960s
- Speed has dramatically increased and cost decreased, especially with VLSI memory chips
- Usually organized in the form of an array in which each cell stores on bit
- Each row stores a memory word and connects to a *word line* for addressing purposes
- Each column connects to a Sense/Write circuit which connects to data input/outline lines of the chip

**Organization of bit cells in a memory Chip**

![Diagram of memory chip organization](image-url)
RAM Memories

- Memory chips organization can be referred to by their internal storage or its total storage
  - 16 words of 8 bits each is: a “16x8 organization
  - or “a 128 x 8” memory chip

- The organization determines the number of external connectors required in a chip
  - power and ground connections are also required

Organization of 1K x 1 Memory Chip
• A 16 x 8 memory chip requires 16 external connections
  – 4 address
  – 8 data
  – R/W & CS
  – Power & ground

• A 1k (1024) memory chip:
  – can be organized as 126 x 8
  – would require 19 external connections
  – alternately can be organized as 1K x 1
  – requires 16 external connections
Types of RAM
(Dynamic RAM)

• DRAM: Dynamic RAM
  – used for most system main memory because it is cheap and small
    • standard RAM, typically rated 60 or 70 ns
  – must be continually rewritten in order for it to maintain its data (otherwise the charge fades)
    • done by placing the memory on a refresh circuit that rewrites the data several hundred times per second.
  – singled ported
    • only one port for accessing data-either writing or reading (can do both simultaneities)

• FPM DRAM: Fast Page Mode DRAM
  – slightly faster than regular DRAM
  – uses a slightly more efficient method of calling data from the memory
  – not used much anymore due to its slow speed, but it is almost universally supported.

• EDO RAM: Extended Data Output RAM
  – same as DRAM, with faster sequential access
    • allows one access to begin while another is being completed
  – generally 5-20% faster than FPM DRAM
  – must be properly supported by the chipset
  – but it is the most common type of memory for most users
Types of RAM 2
(Dynamic RAM 2)

- **BEDO RAM:** Burst Extended Data Output RAM
  - basically EDO DRAM with combined pipelining technology (read data in fast bursts)
  - capable of working with faster bus speeds.
  - Support for the BEDO technology is rather sparse

- **SDRAM:** Synchronous Dynamic RAM
  - its speed is synchronous directly from the clock speed of the entire system.
  - it works at the same speed as the system bus, up to 100MHz
  - although faster, the speed difference may not be noticed due to the fact that the system cache masks it
  - becoming new standard for PC memory

- **RDRAM:** RAMBus DRAM
  - technology still being developed by Intel
    - may prove to surpass SDRAM
  - goal is to get rid of the latency (time taken to access memory)
    - by actually narrowing the bus path
    - and treating the memory bus as a separate communication channel
Types of RAM 3
(Special Purpose RAM)

- **SRAM: Static RAM**
  - maintains its data as long as power is provided
  - does not need to be refreshed
  - much more expensive than DRAM.
  - very fast and typically used for cache

- **SGRAM: Synchronous Graphics RAM**
  - single ported RAM for some graphics cards
  - provides about 5% more bandwidth than the EDO RAM

- **VRAM: Video RAM**
  - typically only used on graphics accelerators
  - dual ported memory (Can read from memory and write to screen at the same time)
  - significant speed performance over its EDO RAM

- **WRAM: Window's RAM**
  - dual ported, static memory
  - faster than VRAM

- **MDRAM: Multi-banked Dynamic RAM**
  - provides more bandwidth than EDO RAM) through bank switching
  - only used on the ET6000 chipset based graphics cards
  - useful for high resolution and color depth graphics
SDRAM Considerations

• SDRAM is the new developing standard and buying SDRAM requires some information to consider

• Speed
  – speed is generally rated in two different ways
    • most common way is the nanosecond rating, e.g. "10 nanosecond", which is the common speed for SDRAM
    • second method is the MHz rating, e.g. "100 Mhz".

  – SDRAM is synchronous, tied to the bus speed of the system
    • unlike older memory that used wait states to compensate for slowness, SDRAM does not use wait states.
    • this means that the memory must be fast enough to work on intended system
    • therefore, 10ns SDRAM should really not be used in systems using more than an 83MHz bus


SDRAM Considerations 2

• 2-clock vs. 4-Clock
  – structurally, 2-clock and 4-clock, SDRAMs are the same, but they are accessed differently
    • A 2-clock SDRAM module is set up so that each clock cycle accesses two chips on the module.
    • A 4-clock SDRAM setup accesses 4 chips per clock cycle
  – check motherboard's documentation to determine which access number is appropriate
    • 4-clock modules are the most common

• Serial Presence Detect
  – some SDRAM modules have a special EEPROM chip on it that holds information about the SDRAM module, such as speed settings.
  – the motherboard queries this chip for information and makes changes in the settings to work with the SDRAM
  – some motherboards require this feature
    • check motherboard's documentation if the board requires it
    • SDRAM won't work without it.
Rambus Random-Access Memory (RDRAM)

- Developed by Rambus, Inc.
  - In 1997, Intel licensed the technology for use on its motherboards
  - An alternative memory architecture called SyncLink DRAM (SLDRAM)

- RDRAM data transfer at up to 600 MB/sec
  - (SDRAM) can deliver data at a maximum speed of about 100 MB/sec
  - A newer version of RDRAM (nDRAM) transfers data at up to 1,600 MB/sec

- RDRAM is used in place of VRAM in some graphics accelerator boards, and in some main memory

- RDRAM employs a narrow, uniform-impedance transmission line, the Rambus Channel, to connect the memory controller to a set of RIMMs (RDRAM modules)
RDRAM

SDRAM Architecture

- Variable length wires, different routes
- 66-133Mhz

Rambus Architecture

- Few wires: all same length & load
  - All run at high speed
  - More uniform routing
- Precision clocking
- 300 - 400 MHz bus clock
RDRAM

- RDRAM memory controller connects to multiple DIMM sockets through a 64-bit wide memory bus operating at 300-400 MHz
  - High bandwidth

- Address and control signals are connected to the DIMM modules using a different topology than for the data bus, resulting in some signals being loaded differently than others
  - Row and Column addresses are transmitted on a shared set of address lines, with the memory controller scheduling this resource when multiple transactions are being serviced
  - Low pin count
**RDAM**

### DIMM Modules

- 1 rank of devices responds to each access
  - All devices respond similarly
  - All devices consume same power
- Single-sided DIMM
  - 4 banks per device => DIMM has 4 banks

### RIMM Modules

- 1-32 devices per RIMM module
  - 1 device responds to each access
  - Devices can be in different power states
  - Different capacities for different market segments
  - Single-device minimum upgrade granularity
  - Module bandwidth same as device bandwidth
- Devices are independent
  - 8 device RIMM, 16 banks each => RIMM has 128 banks
Read Only Memory (ROM)

- ROM is memory that can only be read from but not written to, “non-volatile memory”
  - used in situations where the data must be held permanently
  - Computer BIOS is usually stored in ROM
  - ROM is slower than RAM
    - often ROM data is copied into RAM before use, “shadowing” which increases operating speed.

- PROM: Programmable ROM
  - basically a blank ROM chip that can be written to once

- EPROM: Erasable Programmable ROM
  - like PROM except, that one can erase stored data, allowing it to be rewritten.
  - erase the ROM by shining ultra-violet light onto a sensor atop the ROM chip for a certain amount of time

- EEPROM: Electrically Erasable Programmable ROM “Flash ROM”
  - like EPROM but erased by an electrical current
  - EEPROM allows users to upgrade their BIOS
Memory Packaging

- Packaging is the entire makeup of a unit of memory
  - memory chips are too small, they must be in a medium and combined on a small fiberglass
    - so that can be easily manipulated and added to a system

- DIPs: Dual Inline Packages
  - individual memory units either soldered onto the motherboard or placed in special sockets
    - when a motherboard-soldered memory chip went bad, the motherboard required replacement
    - With socketed chips, chip creep was a problem when chips were lodged out of the socket due to thermal expansion

- SIMMs: Single Inline Memory Modules
  - cards latched into a socket on the motherboard eliminates previous problems
  - come in two sizes, 30-pin and 72-pin
    - 30 pin SIMMs usually came with small amounts of memory (smaller than 8MB); not used much anymore
    - 72-pin SIMMs are mostly used
Memory Packaging 2

- both single sided and double sided designs (whether the SIMM has DIP chips on one side of the SIMM or both)
  - Usually, 1, 4, and 16MB SIMMs are single sided. Other sizes are double sided
- Some double sided SIMMs are actually two single sided SIMMS back to back, wired together within the fiberglass module
  - these designs operate a little different electrically
  - which is the reason some motherboards only use SIMMs of certain sizes

- DIMMs: Double Inline Memory Modules
  - a newer memory module with 168 pins
    - 83 pins on each side of the DIMM are more like little pads of metal
  - provides a 64-bit memory pathway, allowing more performance while maintaining a small package size
  - also makes it even more imperative that the connection remain intact with all the "pads."
  - SDRAM usually comes on DIMMs
  - come in either 3.3 volt or 5 volt designs, and unbuffered or buffered
    - 3.3 volt unbuffered is most common
Static Ram

• Contain circuits that retains their state as long as power is applied

• Implementation
  – cross connect two inverters to form a latch
  – transistors act as switches that open or close under the control of the Word Line

• Operation
  – Write: Sense/ write circuit places value on line b and compliment on b’; forces cell into correct state
  – Read: Activate Word Line to close switches $T_1$ and $T_2$; b carries the value of the circuit; Sense/ write circuit monitors b and b’ and set out accordingly

A Static RAM Cell
CMOS Memory Cell

- Major advantage of very low power consumption
  - current flows only when the cell is being accessed
  - 5 volt and 3.3 volt versions

- Implementation
  - transistor pairs forms the inverters
  - in state 1, point X is high
    - transistors T3 and T6 are on while T4 and T5 are off

A CMOS Memory Cell

\[ \text{b} \quad \begin{array}{c}
T_1 \\
T_3 \\
X \\
T_5 \\
T_6 \\
T_4 \\
Y \\
T_2
\end{array} \quad \text{b}' \]

\[ \text{V}_{\text{supply}} \]

Word Line

Bit Lines

Copyright 1999 ©Roy M. Wnek. All rights reserved.
Static Ram

- Contain circuits that retains their state as long as power is applied

- Implementation
  - cross connect two inverters to form a latch
  - transistors act as switches that open or close under the control of the Word Line

- Operation
  - Write: Sense/ write circuit places value on line b and compliment on b’; forces cell into correct state
  - Read: Activate Word Line to close switches T₁ and T₂; b carries the value of the circuit; Sense/ write circuit monitors b and b’ and set out accordingly

A Static RAM Cell
CMOS Memory Cell

- Major advantage of very low power consumption
  - current flows only when the cell is being accessed
  - 5 volt and 3.3 volt versions

- Implementation
  - transistor pairs forms the inverters
  - in state 1, point X is high
    - transistors T3 and T6 are on while T4 and T5 are off

A CMOS Memory Cell

![Diagram of CMOS Memory Cell]
Dynamic Memories

- Circuits with fewer components translate to cheaper cost

- Information is stored as a charge in a capacitor
  - only able to store a charge for a few milliseconds
  - needs periodic refreshing by a special Refresh Circuit which accesses all cells every 2-16 milliseconds
  - takes relatively long to charge a capacitor

- A transistor acts as a switch in applying a charge to a capacitor

- During a read operation, the cell is discharged and a sense circuit attached to the bit line compares the voltage to a threshold value

A Single Transistor Dynamic Memory Cell
A Typical 1M DRAM Chip

- Organized as a 1k x 1k array

- 10 bits row address and 10 bits column address
  - addresses are multiplexed on 10 pins
  - Row Address Strobe (RAS) and Column Address Strobe (CAS) coordinate addressing

- Block transferring involves multiple column accesses for a single row access

Internal Organization of a 1M x 1 Dynamic Memory Chip
Read Only Memories

- Used to implement parts of a computer memory that contain fixed programs or data
  - BIOS
  - Microprogram Control Store

- Reading from ROM is relatively slow
  - often ROM data is copied to RAM during boot up operation, “Shadowing”

- The presence (or absence) of a connection at point P determines the value of the cell
  - a fuse in used at point P in PROMs
  - a special transistor is used at point P in EPROMS and EEPROMS

A ROM Memory Cell
**Speed, Size, and Cost**

- Relative speed, size, and cost of
  - static RAM
  - dynamic RAM
  - magnetic disk storage

- Basic strategy
  - bring the instructions and data that will be used in the near future as close to the CPU as possible

- Principle of locality and other algorithms

**Memory Hierarchy**

```
Increasing size

CPU
  Primary cache
  Secondary cache
  Main memory
  Disk memory

Increasing speed

Increasing cost per bit
```
Cache Memories

• Effectiveness is based on the Locality of Reference principle
  – sequential execution
  – simple and nested loops

• Temporal and spatial manifestation
  – a recently executed instruction is likely to be executed again very soon
  – instructions in close address proximity to a recently executed instruction are also likely to be executed very soon

• Memory control circuit will fetch from main memory a cache block or cache line for ready use
  – a mapping function specifies a relationship between blocks in cache and those in main memory
  – if the desired block is already in cache, a hit occurs and there is no need to refer to main memory, else a miss or fault occurs
Cache Memories 2

– if the cache is full when new blocks are coming in, a *replacement algorithm* determines which blocks to remove from cache

• After a cache hit occurs during a write operation, two strategies are possible
  – *write through*: simultaneous update of the cache and main memory
    • simple, but may result in unnecessary writes to main memory
  – *write-back or copy-back*: set a *dirty bit* in the appropriate cache block which indicates that main memory needs updating whenever that block leaves the cache
    • also may result in unnecessary writes to main memory, copies entire block back to main memory even if only one word changed

• If a cache miss occurs during a read operation, two strategies are possible
  – read in the entire block from main memory before providing the desired word to the processor
  – provide the desired word to the processor as soon as it is read from main memory, *load-through*
Mapping functions

- **Direct Mapping**: the simplest method
  - map some number of main memory blocks to each cache block position
    - use a block number based function
  - use a tag value to identify which block is in cache (only need to check one tag)
  - can cause contention even when the cache is not full

- **Associative Mapping**: a more flexible method
  - place a main memory block in any cache position
  - use a tag value to identify which block is in cache (must check all tags)
  - can cause contention only when the cache is full

- **Set Associative Mapping**: combination method
  - group cache blocks into sets
  - mapping allows a block of main memory to be reside in any block of a specific set
  - eases contention problem of the direct method by providing a few placement choices
Replacement algorithms

- Cache Coherence problem
  - ensuring that two different entities (the CPU and a DMA subsystem) use the same copies of data
  - a valid bit is used to identify the currency of cache blocks when DMA processes bypass the cache
  - cache-flushing may be necessary before a DMA transfer takes place

- No replacement strategy is needed in Direct Mapping

- In Associative and Set-Associative Mapping, a strategy is needed to determine which cache block to remove
  - strong influence on overall system performance

- Least Recently Used (LRU) is commonly used
  - based on the locality principle
  - cache controller must track references to all blocks via a counter
  - can have poor performance in some circumstances which can be improved by added some randomness to replacement choices
Performance Considerations

• The *price/performance ratio* is a strong indicator of commercial success
  – performance depends on how fast instructions can be brought into the CPU for execution

• Memory Interleaving
  – distribute successive memory address among successive memory modules
  – considerable speed up can be achieved when several memory modules can be busy at one time

• Hit and Miss Ratio
  – excellent indicator of the effectiveness of a particular implementation of memory hierarchy
  – ideally, the entire memory should appear to the CPU as a single memory
  – adverse effect of a miss due to the extra time necessary to fetch the desired data from slower memory
  – average access time = function (hit rate, Cache access, and Miss penalty)

\[ t_{\text{ave}} = hC + (1-h)M \]
Performance Considerations 2

– improve hit ratio by
  • larger cache
  • increase the block size, up to a certain point
– reduce the miss penalty
  • use the load-through approach
  • use a secondary cache

• Caches on the CPU chip
  – size of primary (on-CPU) cache is determined by cost and chip size, layout, heat generation
  – add a larger, but slower secondary (off-CPU) cache

• Write Buffer
  – when using a write-through strategy, the CPU writes to a buffer and resumes processing
  – dump the write buffer whenever memory is available
  – check the write buffer before a read operation

• Prefetching
  – via a software instruction executed while the CPU is engaged in computationally intensive operation
  – a prefetch can be inserted by the programmer or the compiler

• Lockup-Free Cache
  – a cache that support multiple outstanding misses
  – avoid locks from prefetching and from misses
Virtual Memories

- In modern computers, address space used by the CPU is larger than physical main memory.
- A Memory Management Unit (MMU) translates virtual addresses to physical addresses.
- Data not in physical main memory is stored on disk and retrieved when needed via DMA.

Virtual Memory Organization

```
Processor -> MMU
  | Virtual address
  | Physical address
  | Data

Cache
  | Physical address
  | Data

Main memory
  | DMA transfer

Disk storage
```
Virtual memory address translation

- Organize programs and data in fixed length units, called *pages*
- Maintain a page table which correlates virtual page numbers to physical page frames
  - control bit indicate the status of the page

Virtual address from processor → Page table address → Virtual page number | Offset → PAGE TABLE → Control bits | Page frame in memory → Page frame | Offset → Physical address in main memory
Associative-mapped TLB

- Translation Lookup Buffer is a small portion of the page table which is stored within the MMU
  - contains the most recently accessed pages
  - determine whether a hit or miss occurs

Virtual address from processor

Virtual page number

Offset

TLB

Virtual page number | Control bits | Page frame in memory
---------------------|-------------|---------------------

Virtual page number

Control bits

Page frame in memory

Physical address in main memory

Page frame

Offset

Yes

Miss

Hit