CIS 314
Memory Organizations

Prof. Michel A. Kinsky
Our fourth view of computer organization: The modern digital computer has three major functional hardware units: CPU, Main Memory and Input/Output (I/O) Units.
**Processor- Memory Gap**

- Performance gap: CPU (55% each year) vs. DRAM (7% each year)
  - Processor operations take of the order of 1 ns
  - Memory access requires 10s or even 100s of ns
  - Each instruction executed involves at least one memory access

![Graph showing the relative performance of processor and memory over the years (1980-2010). The x-axis represents the calendar year, and the y-axis represents relative performance. The graph shows a significant gap between processor and memory performance, with processor performance improving at a faster rate.]
Memory Organization

- Memory is organized and accessed in ways to hide this gap
Memory Organization

• Memory is organized and accessed in ways to hide this gap

Reg

L1 $

Ln $

Main Memory

Secondary Memory

4-8 bytes (word)

8-32 bytes (block)

1 to 4 blocks

1,024+ bytes (disk sector = page)
• The fastest memories are expensive and thus not very large

<table>
<thead>
<tr>
<th>Capacity</th>
<th>Access Time</th>
<th>Cost (per GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100s B</td>
<td>ns</td>
<td>$\text{Millions}$</td>
</tr>
<tr>
<td>10s KB</td>
<td>few ns</td>
<td>$\text{100s}$ Ks</td>
</tr>
<tr>
<td>MBs</td>
<td>10s ns</td>
<td>$\text{10s}$ Ks</td>
</tr>
<tr>
<td>100s MB</td>
<td>100s ns</td>
<td>$\text{1000s}$</td>
</tr>
<tr>
<td>10s GB</td>
<td>10s ms</td>
<td>$\text{10s}$</td>
</tr>
</tbody>
</table>
• Our fifth view of computer organization
• A fast memory can help bridge the CPU-memory gap
• The fastest memories are expensive and thus not very large
Memory Technology

• Random Access Memory (RAM)
  ‣ Any byte of memory can be accessed without touching the preceding bytes
  ‣ RAM is the most common type of memory found in computers and other digital devices
  ‣ There are two main types of RAM
    • DRAM (Dynamic Random Access Memory)
      ‣ Needs to be “refreshed” regularly (~ every 8 ms)
      ‣ 1% to 2% of the active cycles of the DRAM
      ‣ Used for Main Memory
    • SRAM (Static Random Access Memory)
Memory Technology

• Random Access Memory (RAM)
  ‣ Any byte of memory can be accessed without touching the preceding bytes
  ‣ RAM is the most common type of memory found in computers and other digital devices
  ‣ There are two main types of RAM
    • DRAM (Dynamic Random Access Memory)
    • SRAM (Static Random Access Memory)
      ‣ Content will last until power turned off
      ‣ Low density (6 transistor cells), high power, expensive, fast
      ‣ Used for caches
Memory Technology

- Single-transistor DRAM cell is considerably simpler than SRAM cell
- This leads to dense, high-capacity DRAM memory chips
RAM Organization

- One memory row holds a block of data, so the column address selects the requested bit or word from that block.
- RAS or Row Access Strobe triggering row decoder.
- CAS or Column Access Strobe triggering column selector.
RAM Organization

- One memory row holds a block of data, so the column address selects the requested bit or word from that block.
RAM Organization

• Latency: Time to access one word
  ‣ Access time: time between the request and when the data is available (or written)
  ‣ Cycle time: time between requests
  ‣ Usually cycle time > access time

• Bandwidth: How much data from the memory can be supplied to the processor per unit time
  ‣ Width of the data channel * The rate at which it can be used
DRAM Packaging

• DIMM (Dual Inline Memory Module) contains multiple chips arranged in “ranks”
  ‣ Each rank has clock/control/address signals connected in parallel (sometimes need buffers to drive signals to all chips), and data pins work together to return wide word
  ‣ A modern DIMM usually has one or two ranks (occasionally 4 if high capacity)
DRAM Packaging

• DIMM (Dual Inline Memory Module) contains multiple chips arranged in “ranks”
Disk Memory Basics

- Hard disk drive (HDD), hard disk, hard drive is the dominant secondary storage device in computer systems

Typically 2 - 8 cm
Disk Memory Basics

- Hard disk drive (HDD), hard disk, hard drive is the dominant secondary storage device in computer systems.

1. Head movement from current position to desired cylinder: **Seek time** (0-10s ms)
2. Disk rotation until the desired sector arrives under the head: **Rotational latency** (0-10s ms)
3. Disk rotation until sector has passed under the head: **Data transfer time** (<1 ms)
Memory Organization

On-chip

CPU

Cache

bus

Memory

25 cycles

25 cycles

25 cycles

25 cycles
Memory Organization

On-chip

CPU

Cache

bus

25 cycles

25 cycles

25 cycles

25 cycles

Interleaved Memory

Memory bank 0

Memory bank 1

Memory bank 2

Memory bank 3
Intel Core i7 Organization

**Intel® Core™ i7 architecture**

- 731,000K Transistors (4 Core, 8 Thread)
- 2.93GHz
- Simultaneous Multi-Threading/Threads per core 2
- Up to 4 Cores in Desktop
- Additional Caching Hierarchy
- 4 instructions per clock cycle,
- 15 Stage Pipe, Enhanced Micro and Macro Fusion
- Deeper Buffers
- SIMD Units 3 128 bit Single cycle SSE
- Macrofusion* in both 32-bit and 64-bit modes

**Diagram:**
- Instruction Fetch and Pre Decode
- Instruction Queue
- Decode
- Rename/Allocate
- Reservation Station
- Execution Units
- Retire Unit (RetOrder Buffer)
- 32kB Data Cache
- 32kB Instruction Cache
- 256kB 2nd Level Cache
- L3 and beyond

*VISUAL COSSO GameController

Computer Architecture and Embedded Systems Laboratory (CAES Lab)
Intel Haswell

4th Generation Intel® Core™ Processor Die Map
22nm Haswell Tri-Gate 3-D Transistors

Quad core die shown above
Transistor count: 1.4Billion
Die size: 177mm²

** Cache is shared across all 4 cores and processor graphics
Next Class

- Caching Principles