Chapter 7.1 - Memory

How is the Hierarchy Managed?

• Registers ↔ Cache
  – by compiler (programmer?)

• Cache ↔ Memory
  – by the hardware

• Memory ↔ Disks
  – by the hardware and operating system (virtual memory)
  – by the programmer (files)

Memory Hierarchy Technology

• Random Access:
  – “Random” is good: access time is the same for all locations
  – DRAM: Dynamic Random Access Memory
    • High density, low power, cheap, slow
    • Dynamic: need to be “refreshed” regularly
  – SRAM: Static Random Access Memory
    • Low density, high power, expensive, fast
    • Static: content will last “forever” (until lose power)

• “Non-so-random” Access Technology:
  – Access time varies from location to location and from time to time
  – Examples: Disk, CDROM, DRAM page-mode access

• Sequential Access Technology: access time linear in location (e.g., Tape)

Main Memory Background

• Performance of Main Memory:
  – Latency: Cache Miss Penalty
  • Access Time: time between request and word arrives
  • Cycle Time: time between requests
  – Bandwidth: IO & Large Block Miss Penalty (L2)

• Main Memory is DRAM: Dynamic Random Access Memory
  – Dynamic, needs to be refreshed periodically (8 ms)
  – Addresses divided into 2 halves (Memory as a 2D matrix):
    • RAS or Row Access Strobe
    • CAS or Column Access Strobe

• Cache uses SRAM: Static Random Access Memory
  – No refresh (6 transistors/bit vs. 1 transistor)
    • Size: DRAM/SRAM - 4-8
    • Cost/Cycle time: SRAM/DRAM - 8-16

Random Access Memory (RAM) Technology

• Why do computer designers need to know about RAM technology?
  – Processor performance is usually limited by memory bandwidth
  – As IC densities increase, lots of memory will fit on processor chip
    • Tailor on-chip memory to specific needs
      - Instruction cache
      - Data cache
      - Write buffer

• What makes RAM different from a bunch of flip-flops?
  – Density: RAM is much denser

Static RAM Cell

6-Transistor SRAM Cell

Write:
1. Drive bit lines (bit=1, bit=0)
2. Select row
3. Cell pulls one line low
4. Sense amplifier on column detects difference between bit and bit

Refresh:
1. Just do a dummy read to every cell.

1-Transistor Memory Cell (DRAM)

Write:
1. Drive bit line
2. Select row

Read:
1. Precharge bit line to Vdd/2
2. Select row
3. Cell and bit line share charges
   • Very small voltage changes on the bit line
4. Sense (fancy sense amp)
   • Can detect changes of ~10^6 electrons
5. Write: restore the value

Refresh:
1. Do a dummy read to every cell.
Chapter 7.1 - Memory

Classical DRAM Organization (square)

- Each intersection represents a 1-Tr. DRAM Cell
- Row and Column Address together: Select 1 bit at a time

Chapter 7.1 - Memory

DRAM Performance

- A 60 ns (t\text{RAC}) DRAM can
  - perform a row access only every 110 ns (t\text{RC})
  - perform column access (t\text{CAC}) in 15 ns, but time between column accesses is at least 35 ns (t\text{PC})
  - In practice, external address delays and turning around buses make it 40 to 50 ns.
- These times do not include the time to drive the addresses off the microprocessor, nor the memory controller overhead.
  - Drive parallel DRAMs, external memory controller, bus to turn around, SIMM module, pins…
  - 180 ns to 250 ns latency from processor to memory is good for a “60 ns” (t\text{RAC}) DRAM

Chapter 7.1 - Memory

Something new: Structure of Tunneling Magnetic Junction

- Tunneling Magnetic Junction RAM (TMJ-RAM)
- Speed of SRAM, density of DRAM, non-volatile (no refresh)
- “Spintronics”: combination quantum spin and electronics
- Same technology used in high-density disk-drives

Chapter 7.1 - Memory

Main Memory Performance

- CPU/Max 1 word; Max/Cache, Bus, Memory N words (Alpha: 64 bits & 256 bits)
- Simple:
  - CPU, Cache, Bus, Memory same width (32 bits)
- Interleaved:
  - CPU, Cache, Bus 1 word; Memory N Modules (4 Modules); example is word interleaved

Chapter 7.1 - Memory

Increasing Bandwidth – Interleaving

- Access Pattern without Interleaving:
  - D1 available
  - Start Access for D1
  - Start Access for D2

- Access Pattern with 4-way Interleaving:
  - We can Access Bank 0 again

Chapter 7.1 - Memory
Main Memory Performance

- Timing model
  - 1 to send address,
  - 4 for access time, 10 cycle time, 1 to send data
- Cache Block is 4 words
- Simple M.P. = 4 x (1 + 10 + 1) = 48
- Wide M.P. = 1 + 10 + 1 = 12
- Interleaved M.P. = 1 + 10 + 1 + 3 = 15

<table>
<thead>
<tr>
<th>address</th>
<th>address</th>
<th>address</th>
<th>address</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 4</td>
<td>1 8</td>
<td>2 12</td>
<td>3 13</td>
</tr>
<tr>
<td>Bank 0</td>
<td>Bank 1</td>
<td>Bank 2</td>
<td>Bank 3</td>
</tr>
</tbody>
</table>

Chapter 7.1 - Memory 1