Memory Hierarchy—Main Memory and Enhancing its Performance

Prof. Tsahi Birk and Dr. Freddy Gabbay
EE Department, Technion
RAM Chip Organization

• Organized in the form of a 2-dimensional matrix.
• Each cell stores one bit.
• Each row stores a memory word and connects to a word line for addressing.
• Each column connects to a Sense/Write circuit which connects to data input/output lines of the chip.

Remarks:
• “memory word” = row = \(\sqrt{\text{chip capacity}/k}\) for k-bit-wide chip.
• Replicate entire structure N/k times for N-bit CPU word.
Various Memory Cells

- **SRAM cell:**
  - Cross connect 2 inverters.
  - Pass transistors act like switches for read/write operations.
  - 6 MOS transistor/1-bit cell.

- **DRAM cell:**
  - Information is stored as a charge in a (MOS) capacitor.
  - Charge can be stored for a few milliseconds - needs periodic refreshing by a special Refresh Circuit.
  - A pass transistor acts as a switch in applying a charge to a capacitor.

- **ROM cell:**
  - Used for parts of a computer memory containing fixed programs or data (BIOS, Microprogram, etc.)
  - Reading from ROM is relatively slow (often ROM data is copied to RAM during boot)
  - A fuse in used at point P in PROMs
  - A special transistor is used at point P in EPROMS and EEPROMS
Types of RAM

• **SRAM:**
  – Maintains its data as long as power is provided.
  – does not need to be refreshed
  – more expensive than DRAM.
  – very fast and typically used for cache
  – Address is not multiplexed.

• **DRAM:**
  – Used for most system main memory because it is cheap and dense (bits per mm^2).
  – Requires refreshing circuit.
  – Addresses divided into 2 halves:
    » **RAS** or *Row Access Strobe*
    » **CAS** or *Column Access Strobe*
  – Only one port for accessing data-either writing or reading (can do both simultaneities).
Fast Page Mode RAM

- FPP RAM: Fast Page Mode RAM
  - A FPM DRAM access consists of a sequence of accesses.
  - Only the first access specifies both row and column addresses.
  - Successive accesses to the same row of memory only require the column address.
  - Also, can have one row of SRAM, which is loaded in parallel by RAS; accesses within that row are fast because 1) no RAS and 2) SRAM cells.
Main Memory Performance Parameters

- **Access Time**: time that elapses between the initiation of an operation and arrival of data.
- **Cycle Time**: minimum time delay required between the initiation of two successive memory operations.
Main Memory Performance

- **Simple:**
  - CPU, Cache, Bus, Memory same width (32 bits)

- **Wide:**
  - CPU/Mux 1 word; Mux/Cache, Bus, Memory N words (Alpha: 64 bits & 256 bits)

- **Interleaved:**
  - CPU, Cache, Bus 1 word: Memory N Modules (4 Modules); example is word interleaved

FIGURE 5.31 Three examples of bus width, memory width, and memory interleaving to achieve higher memory bandwidth.
Main Memory Performance

- **Timing model**
  - 1 to send address,
  - 6 access time within chip, 1 to send data
  - Cache Block is 4 words

- **Simple Mem.** = 4 \times (1+6+1) = 32
- **Wide Mem.** = 1 + 6 + 1 = 8
- **Interleaved Mem.** = 1 + 6 + 4 \times 1 = 11
Independent Memory Banks

- Memory banks for independent accesses vs. faster sequential accesses
  - Multiprocessor
  - I/O
  - Miss under Miss, Non-blocking Cache

- **Superbank**: all memory active on one block transfer
- **Bank**: portion within a superbank that is word interleaved

![Diagram showing the relationship of superbanks and banks.](image)
Independent Memory Banks

• How many banks?
  number of banks \( \geq \) number clocks to access word in bank
  – Full benefit only for sequential accesses (mod no. of banks), otherwise will return to original bank before it has next word ready

• Increasing DRAM chip capacity => fewer chips => harder to have banks
  – min chip capacity \( \times \) no. of banks \( > \) budget
  – increase of memory size in large increments
  – solutions:
    » use older technology for minimal configuration
    » Wider chips
Avoiding Bank Conflicts

• Lots of banks
• Even with 128 banks, since 512 is multiple of 128, conflict (example)

• SW:
  – loop interchange
  – declaring (multi-dimensional) array not power of 2

• HW: Prime number of banks
  – bank number = address mod number of banks
  – address within bank = address / number of banks
  – modulo & divide per memory access?
  – address within bank = address mod number words in bank (3, 7, 31)
  – bank number? easy if $2^N$ words per bank
Main Memory Summary

- Wider Memory
- Interleaved Memory: for sequential or independent accesses
- Avoiding bank conflicts: SW & HW
- DRAM specific optimizations: page mode & Specialty DRAM
Appendix

Memory Chip/Module Technologies and Architectures
14

Types of RAM – EDO RAM

- **EDO RAM: Extended Data Output RAM**
  - same as DRAM, with faster sequential access - allows one access to begin while another is being completed
  - EDO DRAM is very similar to FPM (page mode).
  - EDO advantage over FPM: EDO holds the data valid for a longer period than FPM (data is not turned off with CAS, instead it holds the data until the next piece of data is available).
Types of RAM – BEDO RAM

- **Burst Extended Data Output RAM**
  - basically EDO DRAM with combined pipelining technology allowing read data in fast bursts.
  - After the address has been provided, the next “N” data are read in only one clock cycle each.
  - DRAM itself provides address of the next memory location - done by combining EDO DRAM with pipelineing.
  - 2 Important parameters:
    - **Burst length** - actual length of the burst plus the starting address.
    - **Burst type** - determines whether internal address counter will provide sequential ascending page addresses or interleaved page addresses within the burst length.
Types of RAM - SDRAM

- **SDRAM (Synchronous RAM)**
  - Clocked logic serially reads out an entire row or part of it.
  - Burst length is programmable up to the maximum size of the row.
  - Data transfer is synchronous to an input clock:
    - **SDR (Single Data Rate) SDRAM**: Data transferred in the rising clock edge:
      - Clock rate up to 133 MHz.
      - Typical Data BW: 0.8 – 1 GB/sec.
      - Typical latency: 30-90 nsec.
    - **DDR (Double Data Rate) SDRAM**: Data transferred in the rising and falling clock edges:
      - Clock rate up to 166MHz.
      - Typical Data BW: > 2 GB/sec.
      - Typical latency: 18-64 nsec.
  - Very common in PC platforms.
Types of RAM - SDRAM

Typical DDR SDRAM Read operation

- Clock (CK)
- Data strobe (DQS)
- Data (DQ)
- Address

Read starts
Read ends

DQ & DQS are edge aligned to CK/CK

CAS Latency = 2
Types of RAM - SDRAM

Typical DDR SDRAM Read operation

- Clock (CK)
- Data strobe (DQS)
- Data (DQ)
- Address

Read starts
Read ends

DQ & DQS are edge aligned to CK/CK
RAMBUS (RDRAM, DRDRAM)

• Each chip is more like a bank
  – multiple internal banks
  – internal interleaving

• New, higher-level interface:
  – can ask for variable amounts of data
  – split transaction

• Performance (2\textsuperscript{nd} generation):
  – 1.6\text{GB/s} per chip
  – separate buses for row, column and data, so operations can be overlapped
Flash Memory

- Non volatile
- Read speed similar to DRAM
- Writes 10x – 100x slower
- Critically important for battery-operated, diskless appliances.
- Gradually replacing magnetic disk drives.
- New issue: limited number of erasure cycles (endurance)
  - Wear leveling techniques
  - Advanced enhancement techniques