Multi-Core Computer Architecture: Storage and Interconnects Week 5

Aronya Baksy

October 2021

1 Introduction to DRAM

- Basic motherboard architecture: CPU and mounts for heat sinks/fans, slots for peripheral devices, and DRAM slots (dual in-line memory modules, a.k.a DIMMs)
- The North bridge manages communication between CPU and main memory, and is directly connected to the CPU. It is also called the memory controller hub.
- The south bridge manages communication between CPU and I/O devices (e.g.: PCI devices). It is also called the I/O controller hub

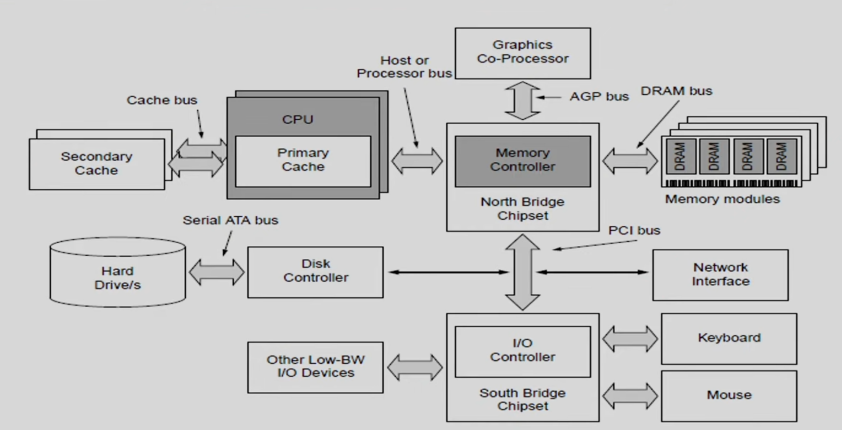

Figure 1: Motherboard Architecture

1.1 Working of SRAM

- When the row is selected, the two transistors are both in the 'on' position.
- On the two sides of the cross-coupled NOT gates, the values are complement of each other. The values are kept into this loop by the cross-coupled not-gates
- For a read operation, it simply involves taking the values on the left side of the cross-coupled NOT gates and placing it into the bit line through the transistor

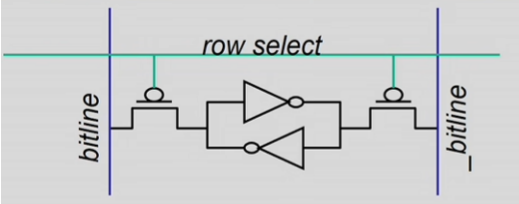

Figure 2: SRAM Cell

- One such cell uses 6 transistors (2 transistors acting as switches, 2 used in each NOT gate)
- For larger memory, these cells are organized into rectangular arrays ($n \times m$ cells). The first n bits of the address select the row (using an $n : 2^n$ decoder) and the next m bits select the column (using an $2^m : 1$ mux)

1.2 DRAM Implementation

- Each bit cell consists of a transistor switch and a single capacitor. Bits are stored as charges on a capacitor.
- Bit cells lose charge on a read, and over time as well the capacitor discharges.
- The sense amplifier amplifies and regenerates the bit-line (to make up for the lost charge when a value is read).
- Comparision of DRAM:
  - Slower access than SRAM because reading requires capacitor discharge which is not under electronic timing
  - High density and lower cost than SRAM
  - Refresh circuitry to keep capacitors from discharging means more power requirement and less efficiency
  - Manufacturing overheads over SRAM (due to capacitor and logic being used together)

1.3 Principle of Interleaving or Banking

- Large monolithic memory chips take longer time to access and don't have provision for multiple parallel accesses
- The memory is divided into banks that can be accessed independently (in the same or in consecutive cycles)
- The address space is partitioned, and bits in the address indicate which bank an address maps to (e.g.: even addresses in bank 0 and odd addresses in bank 1)
- Multiple accesses to the same bank cannot be satisfied. This is called a bank conflict

1.3.1 Page Mode DRAM

- A single bank of DRAM is a 2D array of $m \times n$ cells. A row buffer stores one entire row of this grid when it is to be read
- A memory address is a row and column pair
- A closed row is one that does not exist in the row buffer. Access to a closed row involves the following:
  - **Activate** command places the selected row into the row buffer
  - **Read** or **Write** command that reads/writes the selected column from the row buffer
  - **Precharge** command closes the row and prepares for the next access. Contents of row buffer are stored back to the DRAM row
- Access to an open row (one that exists in the row buffer) does not need the activate command. Rest are same.

1.4 DRAM Hierarchy

- **Channel**: multi-core processors generally have multiple memory controllers. Each memory controller, or each address bus, corresponds to a single channel
- Each channel consists of multiple **DIMMs** (Dual Inline Memory Module) which consist of multiple RAM circuits on a PCB
- Each DIMM has 2 **Ranks** (front side is Rank 0, back is Rank 1). A rank is a set of chips that respond to the same command at the same time but with different pieces of the data

- Each rank consists of **chips**. The rank's data is split into the chips (e.g.: Rank is 64 bits, split into 8 chips)
- Each chip consists of 8 **banks**. The chip's output is selected from one of the banks
- The bank consists of **rows** and columns of bit cells.
- e.g.: Reading 64 byte cache block
  - Each chip has 8 banks. Each bank supplies 1 bit (as per row and col). Hence each chip supplies 8 bits (1 byte)
  - There are 8 chips in 1 rank. So at a time, the rank supplies 8 bytes.
  - Hence, to read 64 bytes, 8 IO cycles are needed. In each cycle the column is incremented
  - Since the entire row is already in the row buffer of DRAM, there is no penalty in the 7 cycles after the first one

2 DRAM Operations

- DRAM controller latency comprises of: queuing, scheduling, converting to basic commands and transfer latency between controller and DRAM
- DRAM bank latency comprises of: Column address strobe (CAS, for open row) or CAS + RAS for closed row), or Pre-charging + CAS + RAS (worst case)
- Parallelism from: multiple banks, or multiple independent channels (this is fully parallel, but more pins, wires, area and power needed)
- Assuming no bank conflicts, data accesses can happen in a pipelined manner with a multi-banked cache

2.1 Address Mapping in Single Channel DRAM

- Parameters: 8 byte memory bus, 2GB memory, 8 chips, 8 banks, 16K rows and 2K columns per bank
- In **row interleaving**, consecutive rows of main memory are in consecutive banks. This allows access to consecutive cache blocks to happen in pipelined manner
- row interleaved address map: 14 bits for row + 3 bits bank + 11 bits column + 3 bits for byte in Bus
- In **cache interleaving**, consecutive cache block addresses in consecutive banks. This allows access to consecutive cache blocks in parallel
- Cache interleaved addr map: 14 bits for row + 8 bits column (high) + 3 bits bank + 3 bits column (low) + 3 bits for byte in bus
- OS performs page colouring to avoid bank conflicts and minimize inter-application interference
- Functions of DRAM controller:
  - Correct operation of DRAM: timing, refresh, address mapping
  - Service DRAM requests while respecting the timing needs
  - Translate requests to DRAM commands
  - Buffer and schedule requests in optimal manner
  - Manage power thermals in DRAM
- DRAM controller can be placed on the chipset (more flexible for different DRAM types, and less power density on CPU) or directly on the CPU (reduced latency, higher bandwidth)
- Logical flow: Schedule transactions (device level), address translation, command scheduling (DRAM bank level), electrical signalling, DRAM access

2.2 DRAM Scheduling Policies

- FCFS
- FR-FCFS: first service requests for already open row, between these use the FCFS. goal is to maximize the row buffer hit rate
- Schedule at the command level. Prefer column commands (read/write) over row commands (activate, pre-charge)
- Prioritize commands based on age, row buffer hit, request type(prefetch/read/write), request mode (load/store), origin of request, criticality, interference to other rows

2.3 Row Buffer Management Policy

- Open row policy: keep row in row buffer after access.
- Closed row policy: if there are no requests in buffer that need the row in the buffer, then close that row
- In case of a row hit, both will have same latency (in closed row it depends)
- In case of row miss, open row policy involves precharge + activate new row + read
- In case of row miss, closed row policy involves activate next row + read + precharge
- Adaptive policy: predict need for a different row or same row and use closed/open row policies appropriately (based on the profile of the application)

2.4 DRAM Refresh

- Memory controller reads each row periodically to refresh the circuits (because capacitors discharge over time)
- this interval is typically 64 ms
- This impacts performance as the DRAM bank is unavailable while being refreshed.
- Burst refresh: all rows refreshed at once, in sequence
- Distributed refresh: each row refreshed at different time, at regular intervals. This eliminates long pauses in DRAM availability

2.5 Address Translation

- CPU translates virtual memory address (logical address) to the physical address via the page table
- Page fault is a situation where the requested virtual address does not have a valid mapping to a physical address in memory. In this case the requested page has to be loaded from disk to RAM and page table updated. OS controls this mechanism
- Servicing page fault:
  - Processor issues I/O signal to disk to read blocks
  - DMA controller issues commands to the disk device to read the data.
  - The DMA controller issues an interrupt once it is done. Processor resumes the interrupted process
- TLB is a cache that stores a subset of page table entries. It reduces AMAT