Memory Hierarchy

Tassadaq Hussain
Riphah International University
Barcelona Supercomputing Center
Universitat Politècnica de Catalunya

Random-Access Memory (RAM)

Key features

- RAM is packaged as a chip.
- Basic storage unit is a cell (one bit per cell).
- Multiple RAM chips form a memory.

Static RAM (SRAM)

- Each cell stores bit with a six-transistor circuit.
- Retains value indefinitely, as long as it is kept powered.
- Relatively insensitive to disturbances such as electrical noise.
- Faster and more expensive than DRAM.

Dynamic RAM (DRAM)

- Each cell stores bit with a capacitor and transistor.
- Value must be refreshed every 10-100 ms.
- Sensitive to disturbances.
- Slower and cheaper than SRAM.

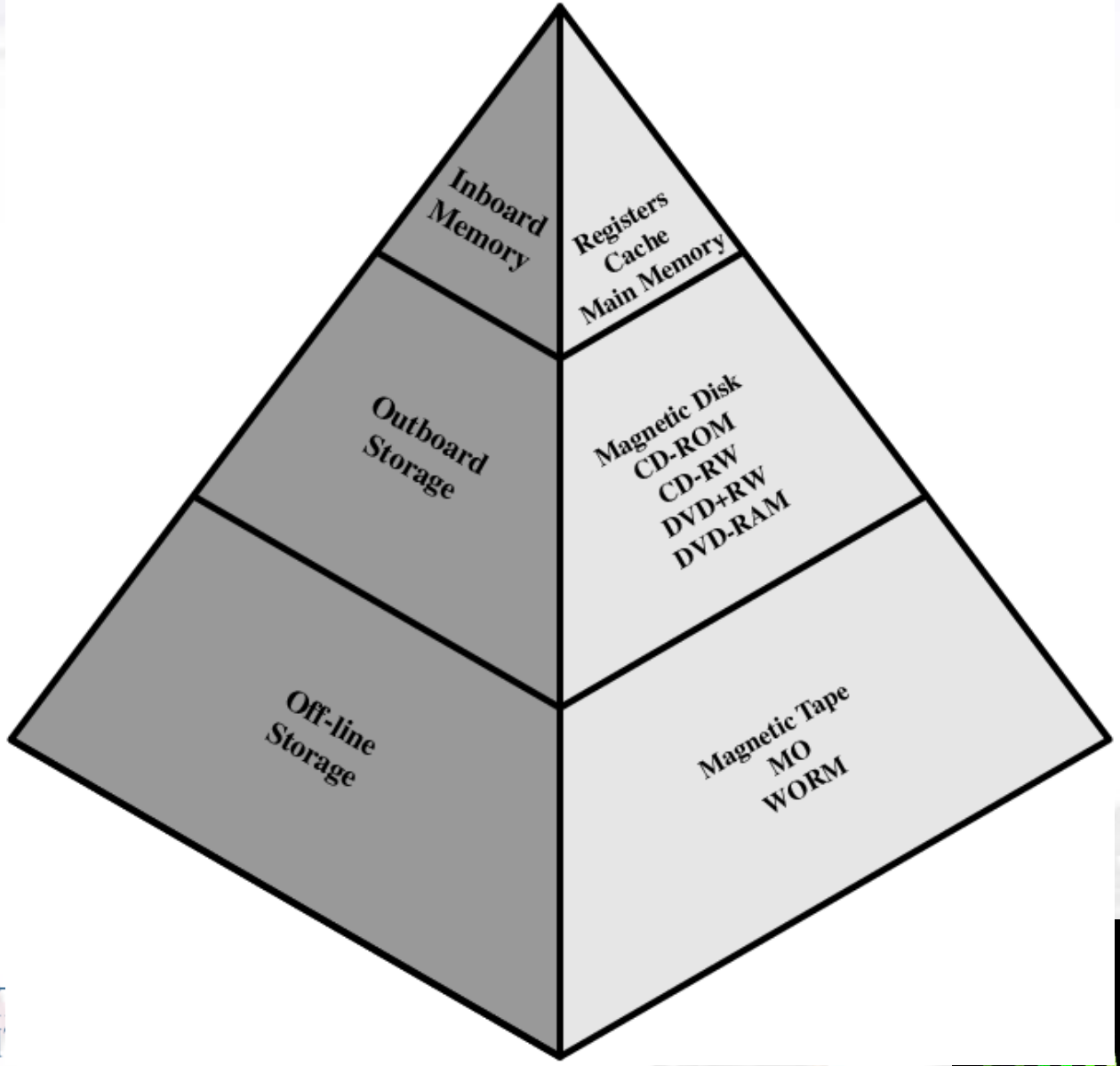
SRAM vs DRAM Summary

	Tran. per bit	Access time	Persist?	Sensitive?	Cost	Applications
SRAM	6	1X	Yes	No	100x	cache memories
DRAM	1	10X	No	Yes	1X	Main memories, frame buffers

Memory Hierarchy

- Registers
  - In CPU
- Internal or Main memory
  - May include one or more levels of cache
  - “RAM”
- External memory
  - Backing store

Memory Hierarchy - Diagram



Characteristics

- Location
- Capacity
- Unit of transfer
- Access method
- Performance
- Physical type
- Physical characteristics
- Organisation

Location

- CPU
- Internal
- External

Capacity

- Word size
  - The natural unit of organisation
- Number of words
  - or Bytes

Unit of Transfer

- Internal
  - Usually governed by data bus width
- External
  - Usually a block which is much larger than a word
- Addressable unit
  - Smallest location which can be uniquely addressed

Access Methods (1)

- Sequential
  - Start at the beginning and read through in order
  - Access time depends on location of data and previous location
  - e.g. tape
- Direct
  - Individual blocks have unique address
  - Access is by jumping to vicinity plus sequential search
  - Access time depends on location and previous location
  - e.g. disk

Access Methods (2)

- Random
  - Individual addresses identify locations exactly
  - Access time is independent of location or previous access
  - e.g. RAM
- Associative
  - Data is located by a mechanism based on placement
  - Access time is independent of location or previous access
  - e.g. cache

Performance

- Access time
  - Time between presenting the address and getting the valid data
- Transfer Rate
  - Rate at which data can be moved

Physical Types

- Semiconductor
  - RAM
- Magnetic
  - Disk & Tape
- Optical
  - CD & DVD
- Others
  - Bubble
  - Hologram

Physical Characteristics

- Switching
- Decay
- Volatility
- Erasable
- Power consumption

Organisation

- Physical arrangement of bits into words
- Not always obvious
- e.g. interleaved

The Bottom Line

- How much?
  - Capacity
- How fast?
  - Time is money
- How expensive?

Hierarchy List

- Registers
- L1 Cache
- L2 Cache
- Main memory
- Disk cache

So you want fast?

- It is possible to build a computer which uses only static RAM (see later)
- This would be very fast
- This would need no cache
  - How can you cache cache?
- This would cost a very large amount

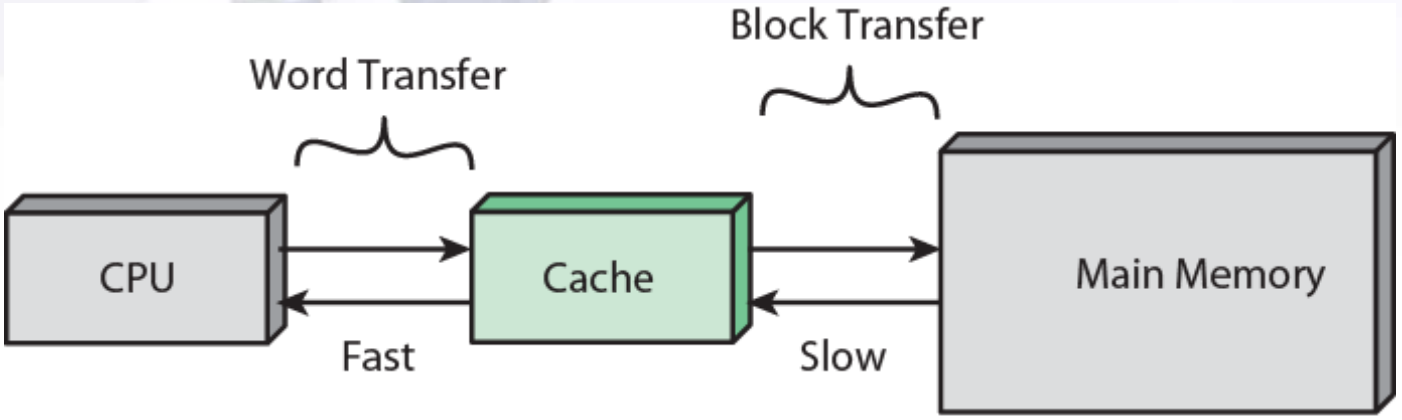
Local Memory System

- Cache
- Scratchpad

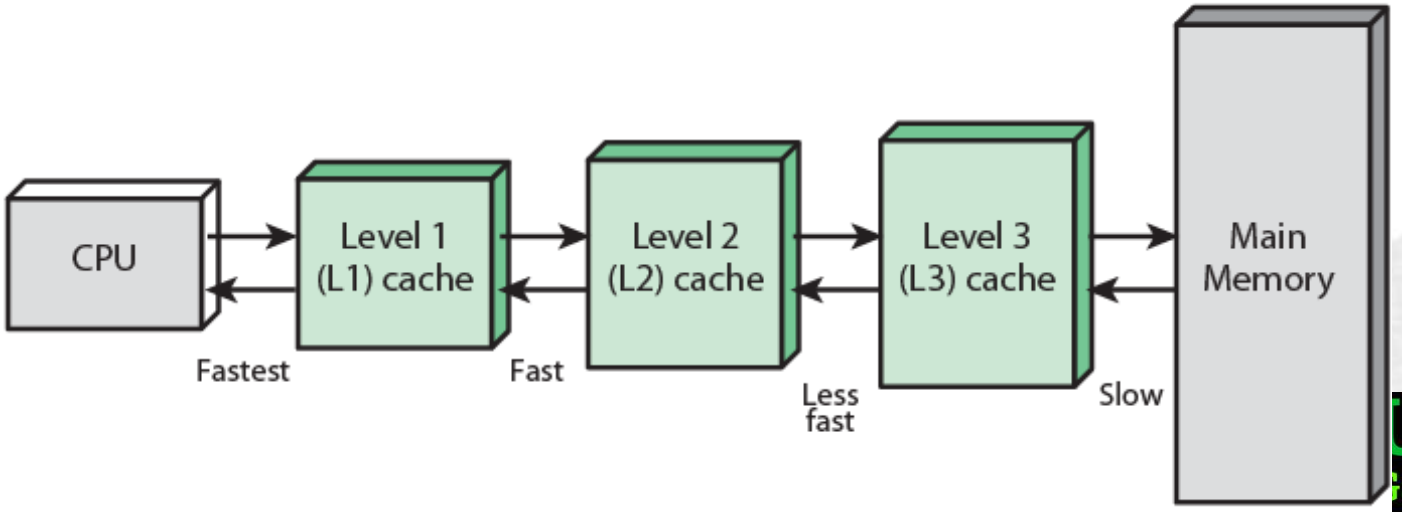
Cache

- Small amount of fast memory
- Sits between normal main memory and CPU
- May be located on CPU chip or module

Cache and Main Memory

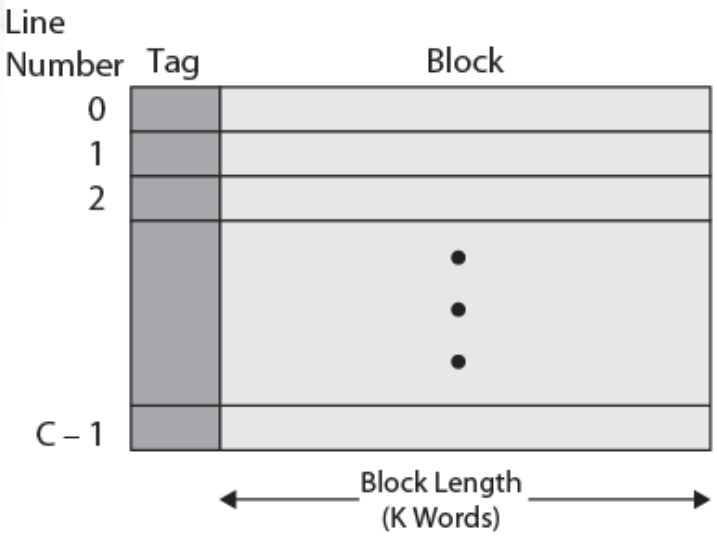


(a) Single cache

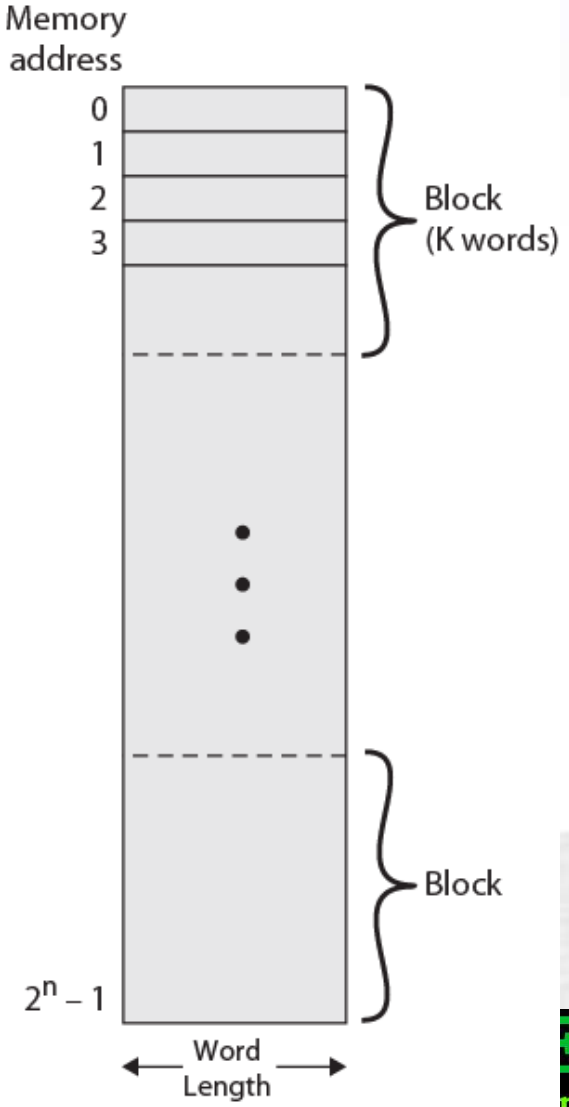


(b) Three-level cache organization

Cache/Main Memory Structure



(a) Cache

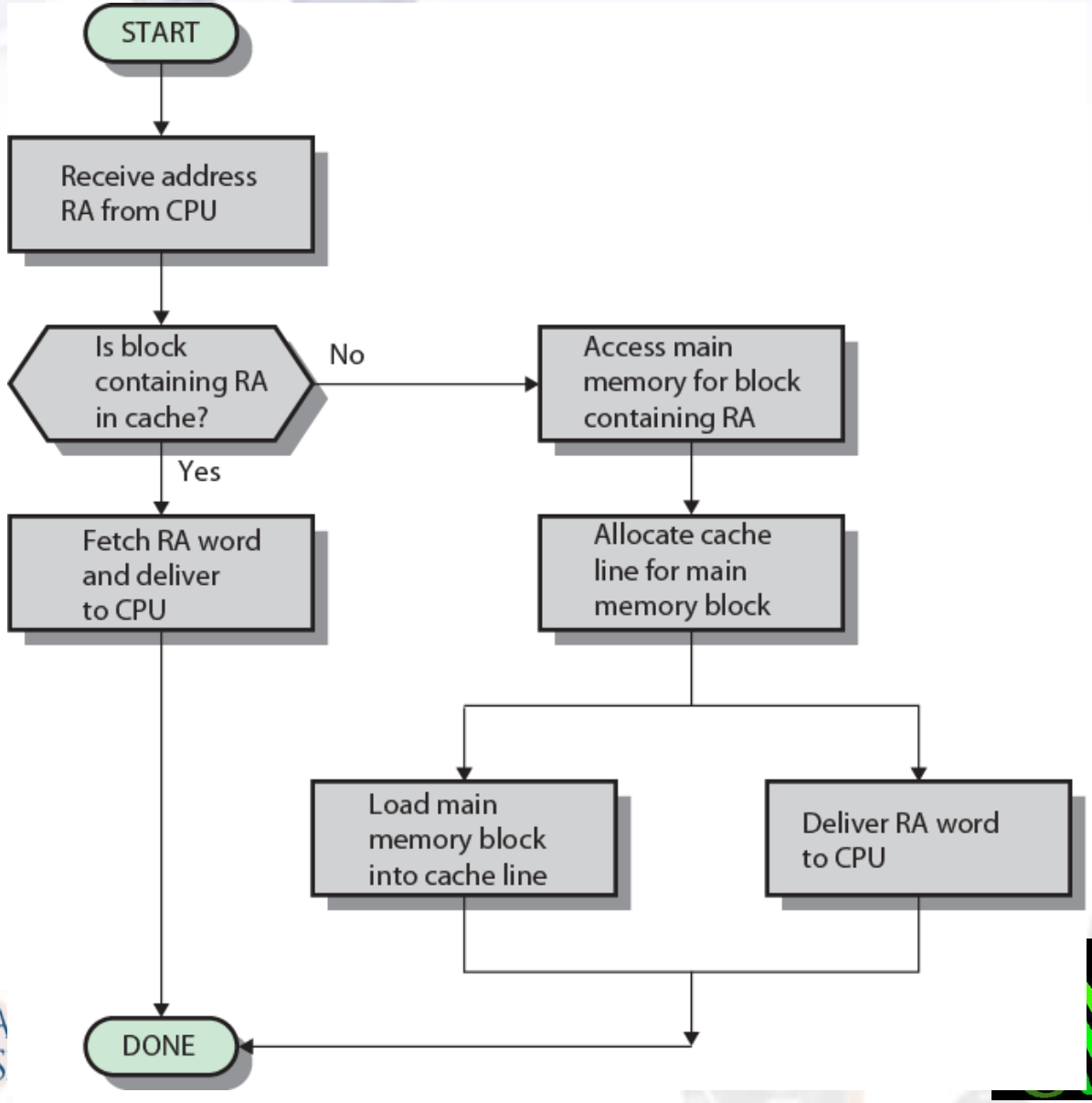


(b) Main memory

Cache operation – overview

- CPU requests contents of memory location
- Check cache for this data
- If present, get from cache (fast)
- If not present, read required block from main memory to cache
- Then deliver from cache to CPU
- Cache includes tags to identify which block of main memory is in each cache slot

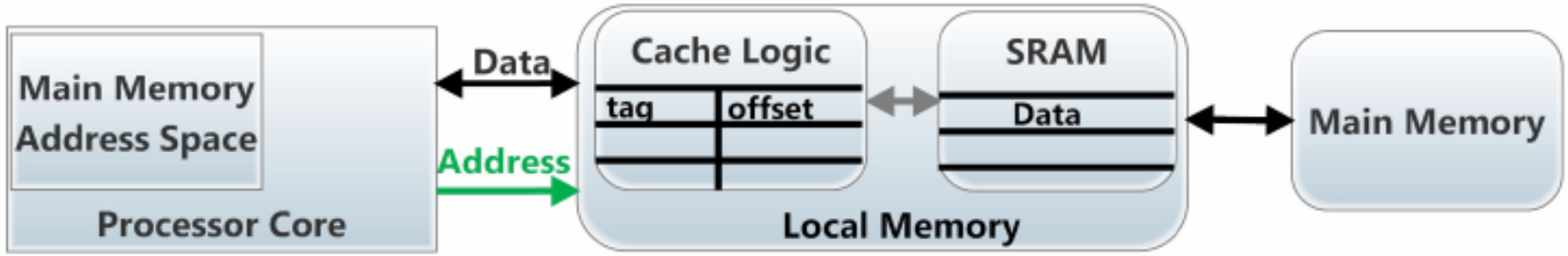
Cache Read Operation - Flowchart



Cache Design

- Addressing
- Size
- Mapping Function
- Replacement Algorithm
- Write Policy
- Block Size
- Number of Caches

A Conventional Memory System Architecture



Cache

- Caches are present in most memory systems.
- The Cache dynamically stores a subset of the frequently used data. Thus, the timing of a load or store operation depends on the relationship between its effective address and the effective addresses of earlier operations.

- Conventional Cache used byte addressable memory.
- 2^b = size of a Cache
- 2^B = size of Main Memory
- Addr = Address of Main Memory
- CL = Data Transfer from/to the memory
- NCL = Number of Cache lines (Cls)
- CLS = Cache Line Size

Direct-Mapped Cache

- Direct Mapped cache is an array of fixed size blocks.
- Each block holds consecutive bytes of main memory data.

Fully associative cache

- A fully associative cache.
- A fully associative cache permits data to be stored in any cache block, instead of forcing each memory address into one particular block. — When data is fetched from memory, it can be placed in any unused block of the cache.

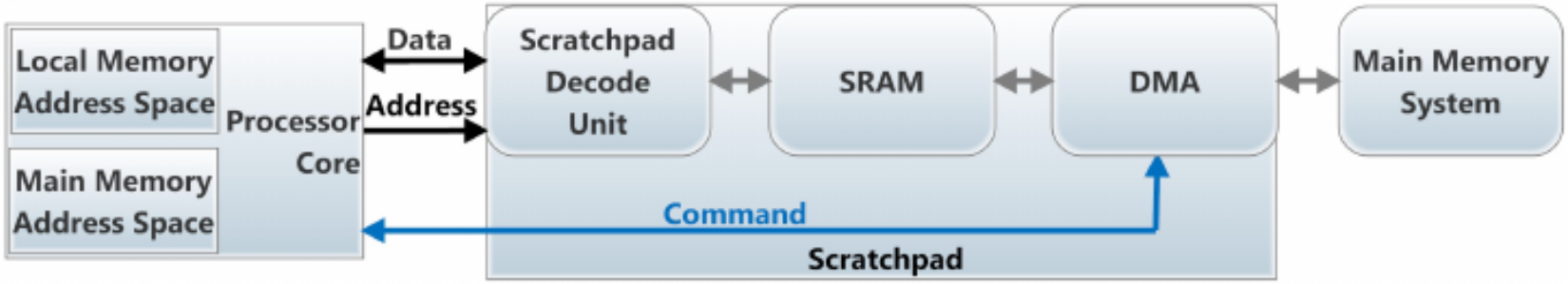
Assignment

- You have 16 bit of address bus
- The maximum size of main memory 64K byte (byte addressable)
- Design a cache having NCL = 8 and CLS = 4K
  - Find out required memory for cache

Set Associative Cache

- A set-associative scheme is a hybrid between a fully associative cache, and direct mapped cache.
- It's considered a reasonable compromise between the complex hardware needed for fully associative caches (which requires parallel searches of all slots), and the simplistic direct-mapped scheme, which may cause collisions of addresses to the same slot (similar to collisions in a hash table).

Scratchpad

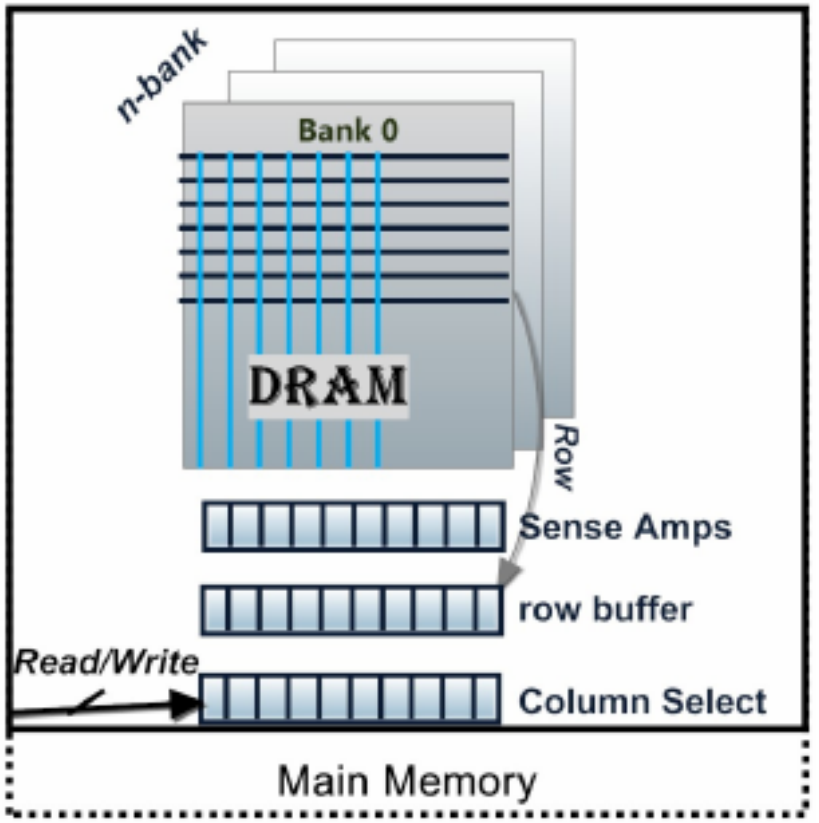


Scratchpad

- The Scratchpad is a fast directly addressed software managed SRAM memory.
- The Scratchpad has better real-time guarantees than caches and by its significantly lower overheads it is better in access time, energy consumption and area.
- Instead of using traditional load/store instructions the scratchpad uses direct memory-memory operations using DMA.
- The Scratchpad memory access uses source and destination address registers, each of which holds a starting address of the memory.

Main Memory System

- DRAM is combination Row x Column
- Row Address
- Column Address



Virtual Memory

- Virtual memory is a memory management capability of an OS that uses hardware and software to allow a computer to compensate for physical memory shortages by temporarily transferring data from random access memory (RAM) to disk storage

- The operating system, using a combination of hardware and software, maps memory addresses used by a program, called virtual addresses, into physical addresses in computer memory.
- Main storage, as seen by a process or task, appears as a contiguous address space or collection of contiguous segments.