COMPUTER ORGANIZATION (IS F242)

LECT 48: PIPELINING

Dynamic Multiple Issue

- Also known as “Superscalar” processors
  - Is an advanced pipelining technology that enables the processor to execute more than one instruction per clock cycle.
  - Instructions issue in-order & processor decides whether 0, 1 or more instructions can issue in the given clock cycle
    - Avoiding structural and data hazards
  - Avoids the need for compiler scheduling
    - Though it may still help
    - Compiled code will always run correctly independent of the issue rate or pipeline structure of the processor

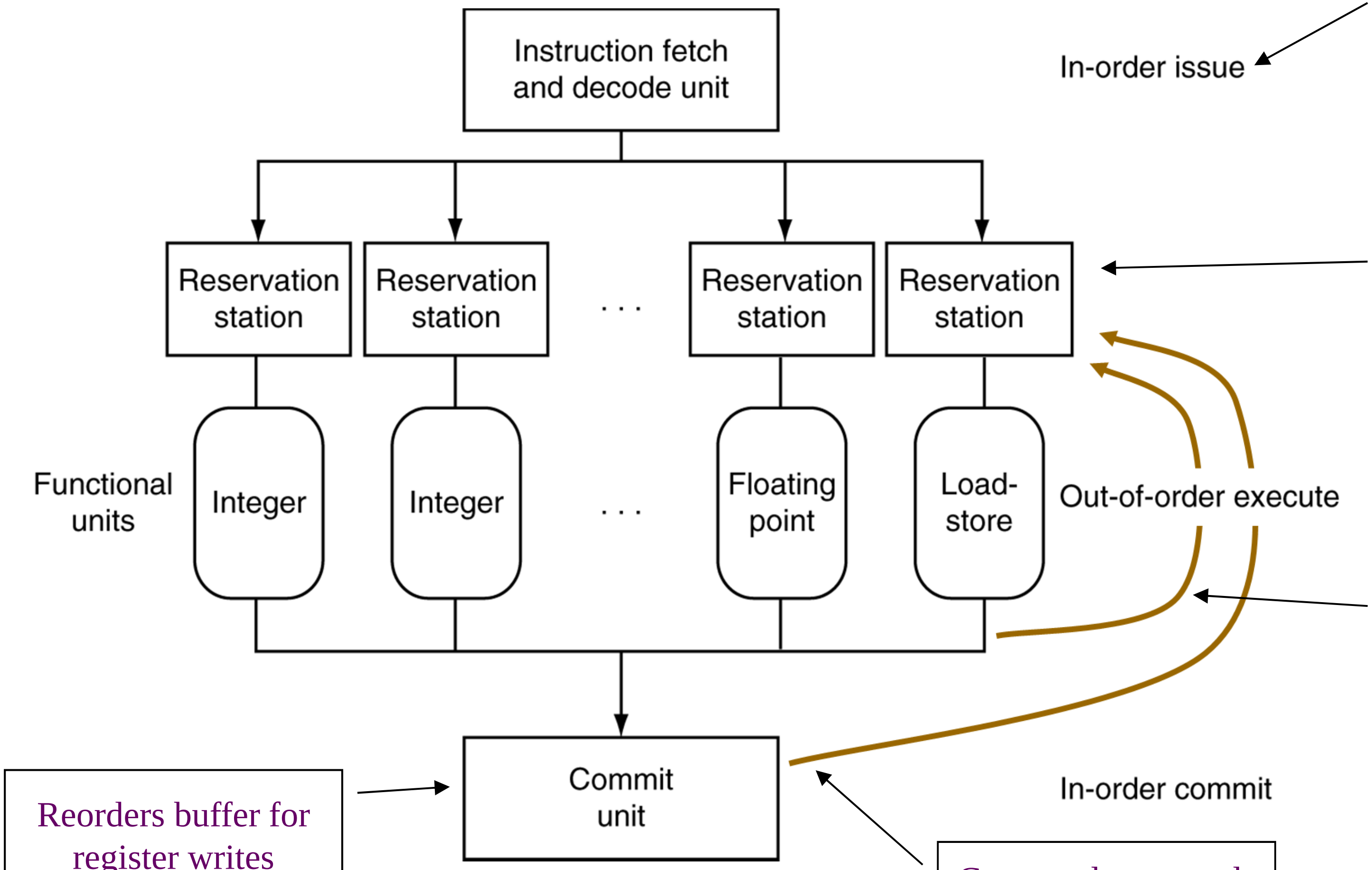
Dynamic Pipeline Scheduling

- Allow the CPU to execute instructions out of order to avoid stalls
  - But commit result to registers in order
- Example

```
lw    $t0, 20($s2)
addu  $t1, $t0, $t2
sub   $s4, $s4, $t3
slti  $t5, $s4, 20
```

  - Can start sub while addu is waiting for lw

Dynamically Scheduled CPU



Instruction fetch and decode unit

Reservation station

Reservation station

. . .

Reservation station

Reservation station

Functional units

Integer

Integer

. . .

Floating point

Load-store

Commit unit

In-order issue

Out-of-order execute

In-order commit

Preserves dependencies

Hold pending operands

Results also sent to any waiting reservation stations

Reorders buffer for register writes

Can supply operands for issued instructions

- Commit Unit
  - decides when it is safe to release the result of an operation to programmer visible registers and memory
- Reservation station
  - A buffer with in a functional unit that holds the operands and the operation
- Reorder buffer
  - holds results in a dynamically scheduled processor until it is safe to store the results to memory or a register

- In order commit
  - A commit in which the results of the pipelined execution are written to the programmer visible state in the same order that instructions are fetched
- Out of order execution
  - A situation in pipelined execution when an instruction blocked from executing does not cause the following instructions to wait

Register Renaming

- Reservation stations and reorder buffer effectively provide register renaming
- On instruction issue to reservation station
  - If operand is available in register file or reorder buffer
    - Copied to reservation station
    - No longer required in the register; can be overwritten
  - If operand is not yet available
    - It will be provided to the reservation station by a function unit
    - Register update may not be required

Speculation

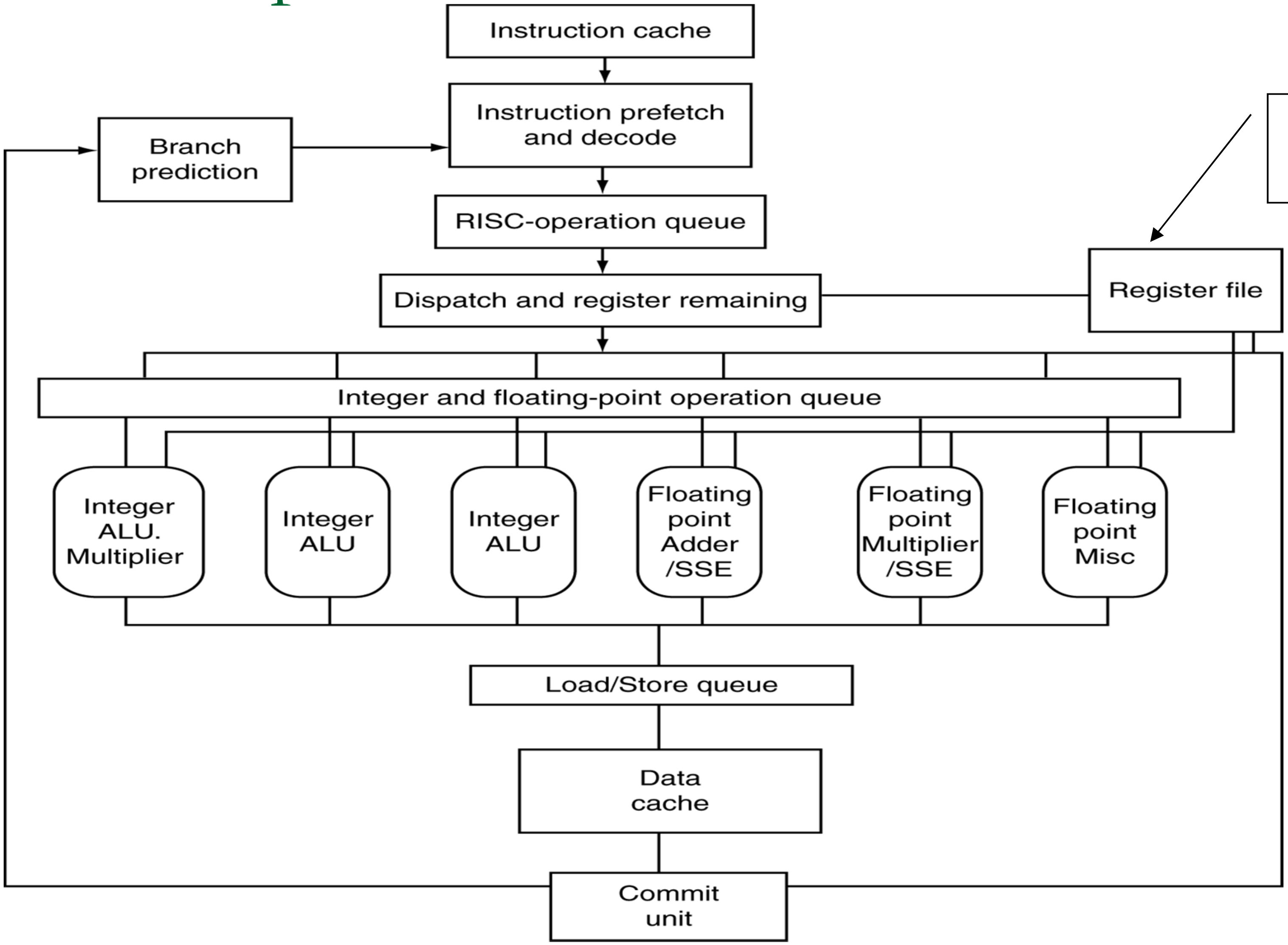
- Predict branch and continue issuing
  - Don’t commit until branch outcome determined
- Load speculation
  - Avoid load and cache miss delay
    - Predict the effective address
    - Predict loaded value
    - Load before completing outstanding stores
    - Bypass stored values to load unit
  - Don’t commit load until speculation cleared

Power Efficiency

- Complexity of dynamic scheduling and speculations requires power
- Multiple simpler cores may be better

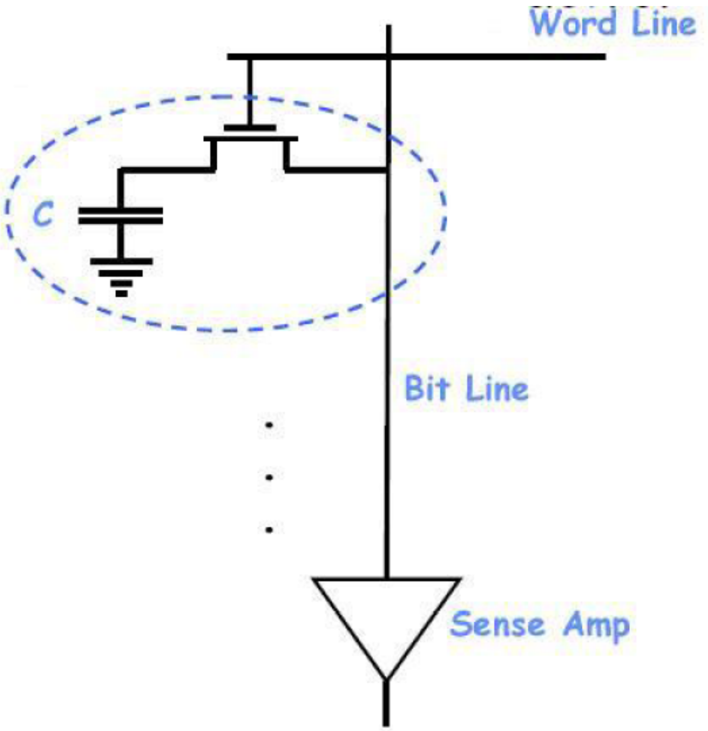
Microprocessor	Year	Clock Rate	Pipeline Stages	Issue width	Out-of-order/ Speculation	Cores	Power
i486	1989	25MHz	5	1	No	1	5W
Pentium	1993	66MHz	5	2	No	1	10W
Pentium Pro	1997	200MHz	10	3	Yes	1	29W
P4 Willamette	2001	2000MHz	22	3	Yes	1	75W
P4 Prescott	2004	3600MHz	31	3	Yes	1	103W
Core	2006	2930MHz	14	4	Yes	2	75W
UltraSparc III	2003	1950MHz	14	4	No	1	90W
UltraSparc T1	2005	1200MHz	6	1	No	8	70W

The Opteron X4 Microarchitecture



72 physical registers

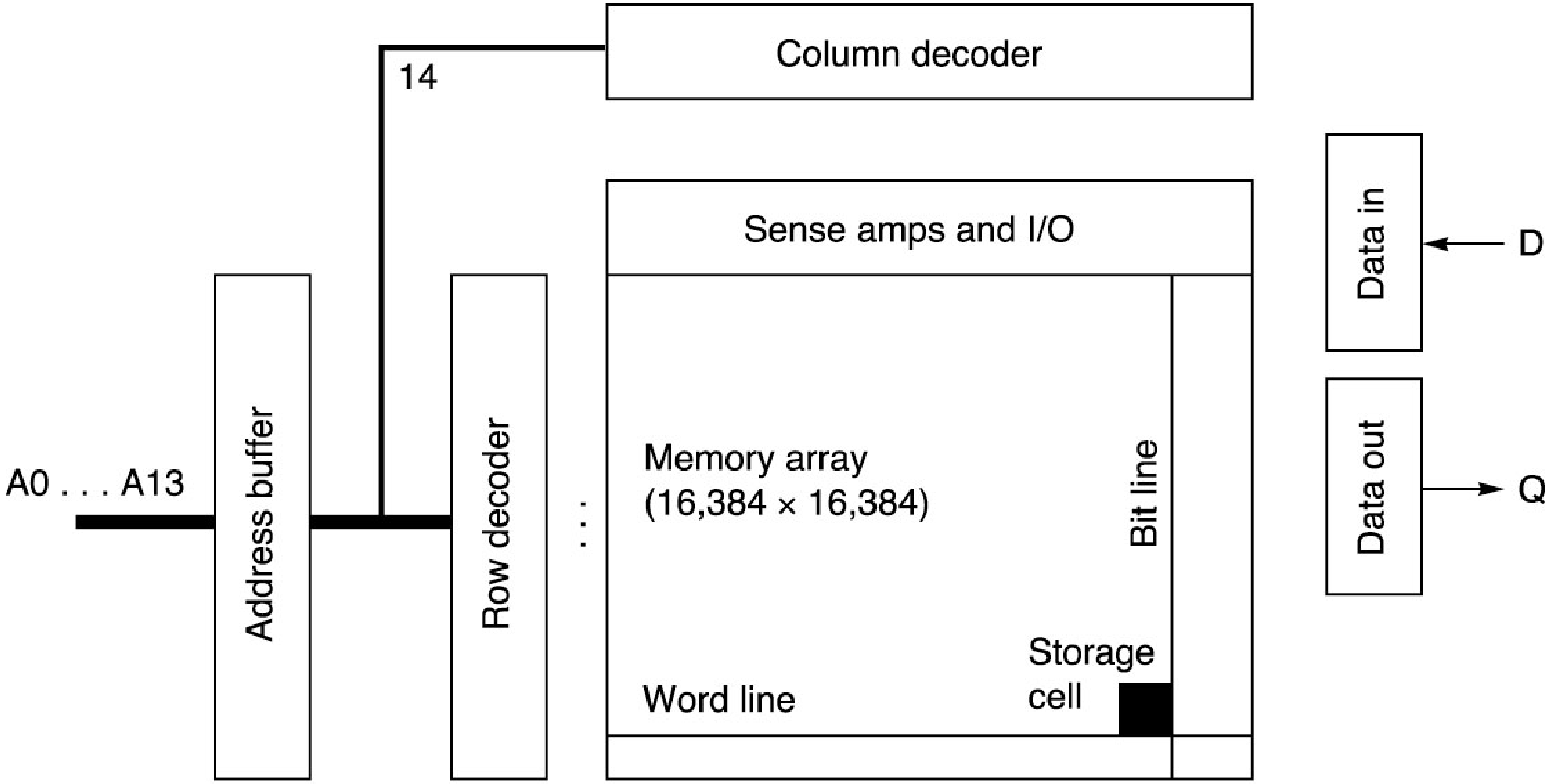
DRAM



Dynamic RAM

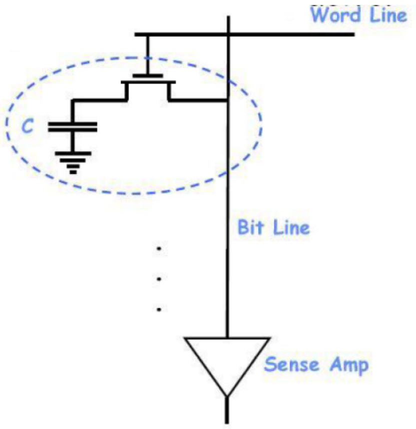
- Bits stored as charge in capacitors(30X 10^-15)
- 1 capacitor + 1 transistor per bit
  - Simpler construction, Smaller per bit , Less expensive
- Used mainly in Main memory
- Charges leak - Need refresh circuits
  - Need refreshing even when powered (4 – 8 ms)
- Sensitive to disturbances
- Slower
- Density (25-50):1 to SRAM
- Address multiplexed

DRAM Organization

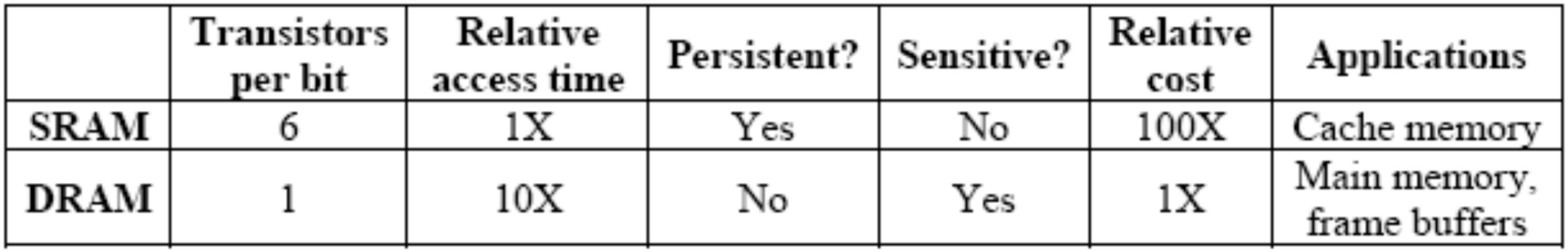


DRAM Operations

- Write
  - Charge bitline HIGH or LOW and set wordline HIGH
- Read
  - Bit line is precharged to a voltage halfway between HIGH and LOW, and then the word line is set HIGH.
  - Depending on the charge in the cap, the precharged bit line is pulled slightly higher or lower
  - Sense amplifier detect change

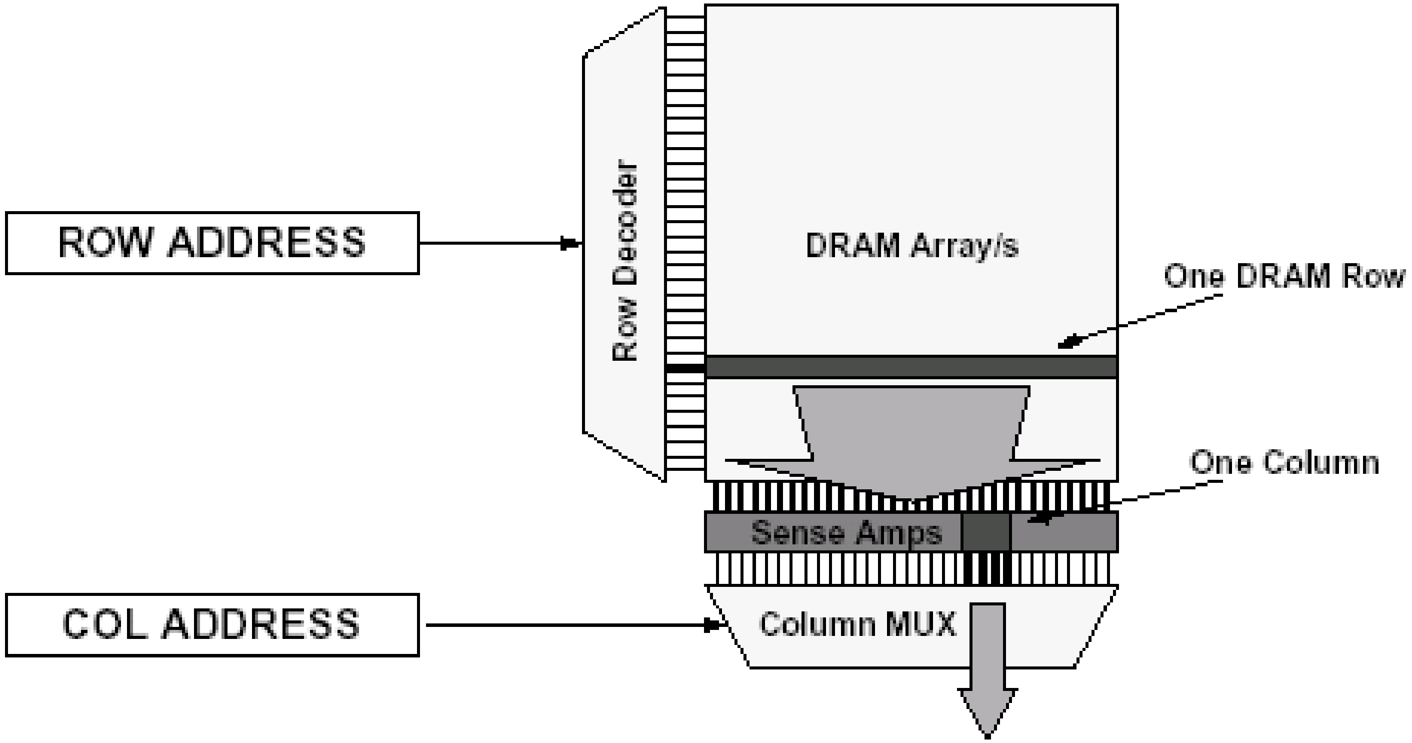


SRAM Vs DRAM



	Transistors per bit	Relative access time	Persistent?	Sensitive?	Relative cost	Applications
SRAM	6	1X	Yes	No	100X	Cache memory
DRAM	1	10X	No	Yes	1X	Main memory, frame buffers

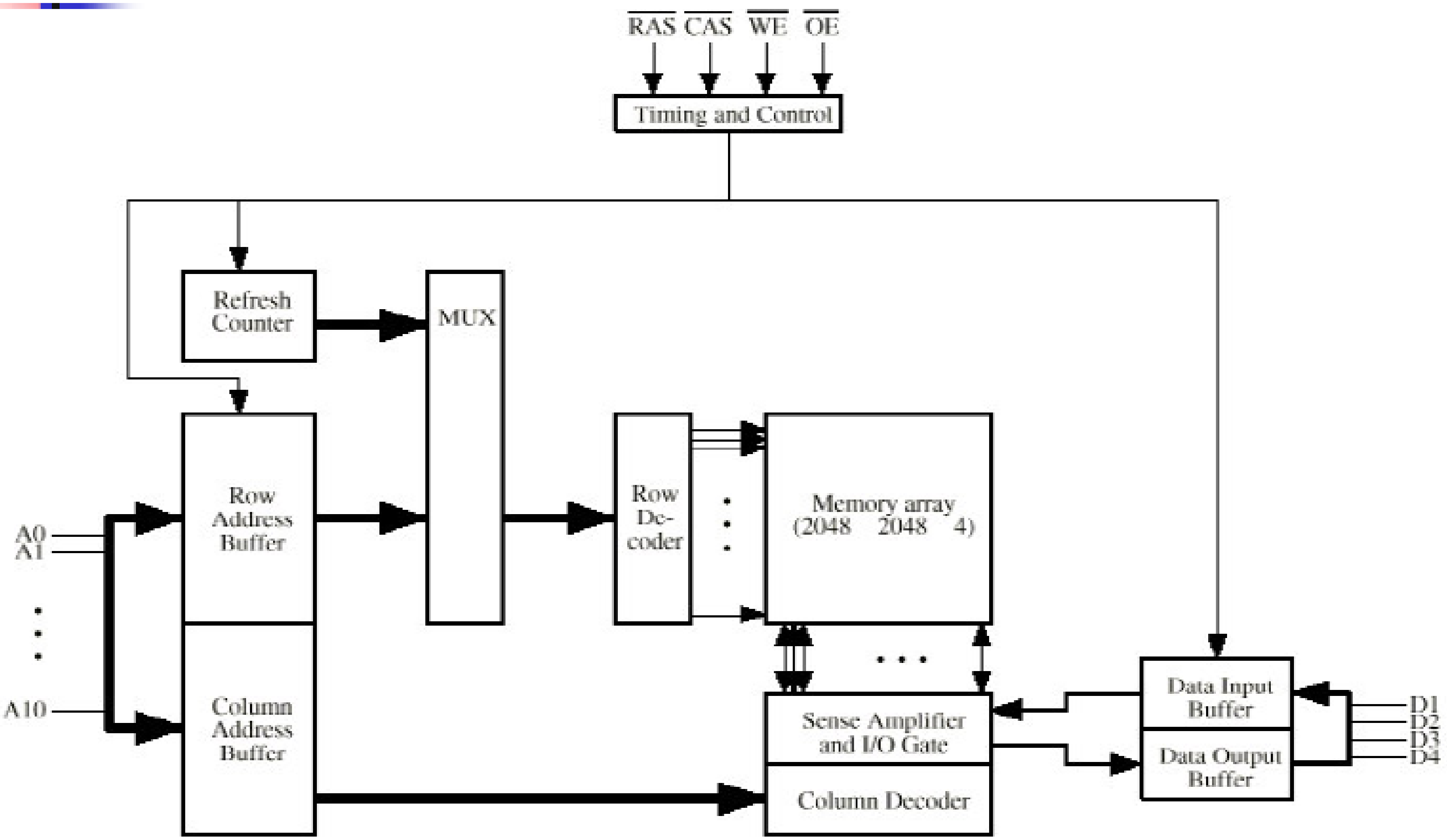
DRAM Access


ROW ADDRESS
Row Decoder
DRAM Array/s
One DRAM Row
One Column
Sense Amps
COL ADDRESS
Column MUX

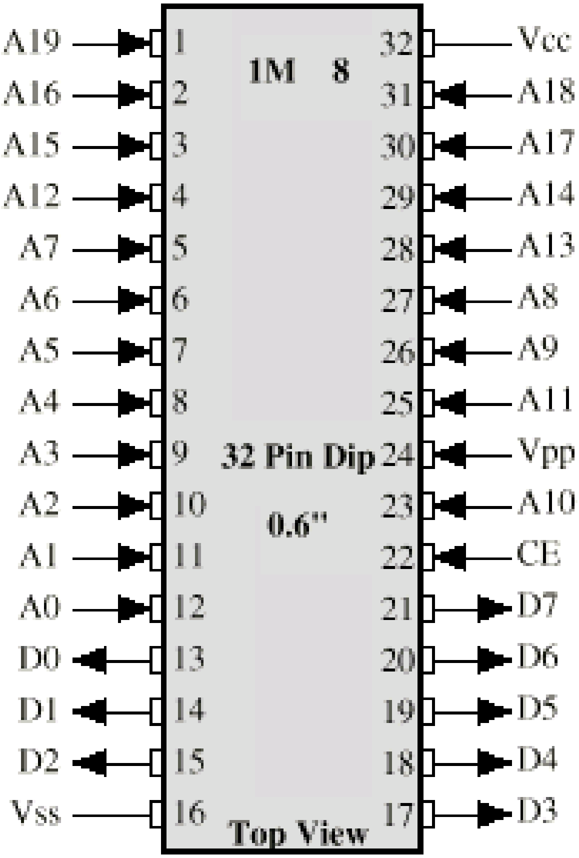
Organization in detail

- A 16Mbit chip can be organized as 1M of 16 bit words
- A bit per chip system has 16 lots of 1Mbit chip with bit 1 of each word in chip 1, bit 2 of each word in chip 2 and so on…..
- A 16Mbit chip can be organized as a 2048 X 2048 X 4 bit array
    - Reduces number of address pins
        - Multiplex row address and column address
        - 11 pins to address (2^11 = 2048)
        - Adding one more pin doubles range of values so X4 capacity

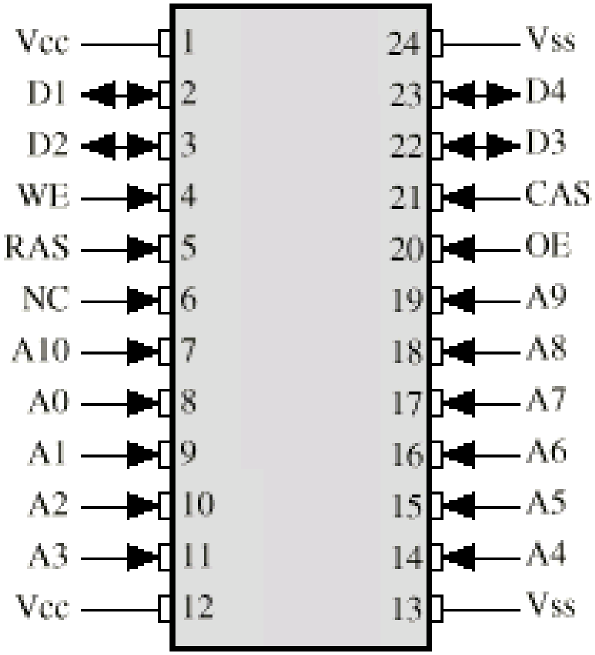
Typical 16M DRAM (4M X 4)



Packaging

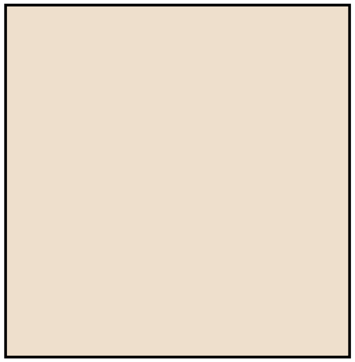


(a) 8 Mbit EPROM



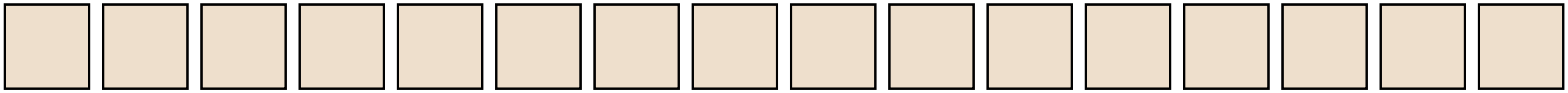
(b) 16 Mbit DRAM

Conventional DRAM Architectures



16 Mb (16M×1) chip

One 4096×4096 array of data bits



16 Mb (1M×16) chip

16 1024×1024 arrays of data bits

- Interface is either the original asynchronous interface or one of the many recent minor modifications of it
- RAS: Row Address Strobe (Send row address when RAS asserted)
- CAS: Column Address Strobe (Send column address when CAS asserted)
- DRAM asynchronously controlled by processor

Memory cells

- D- Latch
  - Datain, Dataout, Select, R/W
- Address Space
- Addressability
- Organization
  - 1-D
  - 2-D
  - 2.5 D