Processors and Memory

System Processors

CPU Architecture

Memory Systems

Memory Addressing

Memory Systems and Technologies

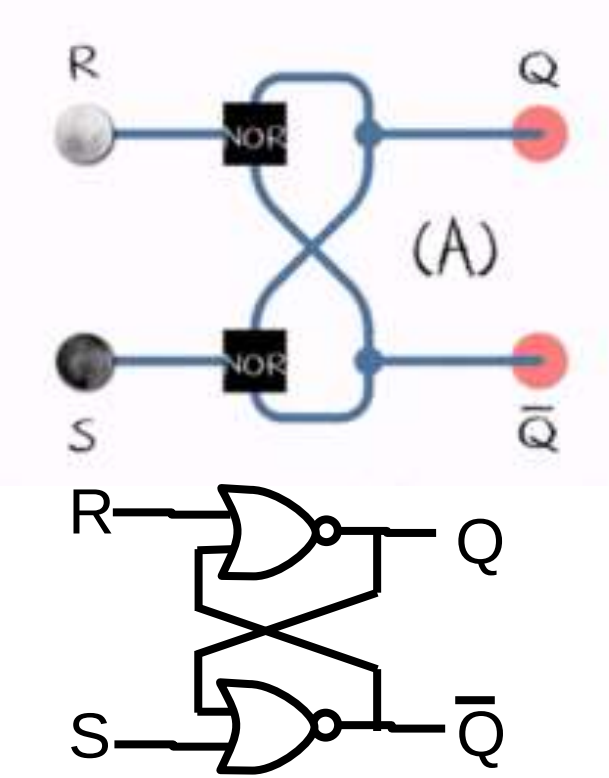
- Two main types of memory:

- Static
  - expensive but fast
  - doesn't need refreshing
  - used by cache

- Dynamic
  - cheaper, higher density of storage, but slower requires refreshing or the information is lost used by main memory

Dynamic Memory

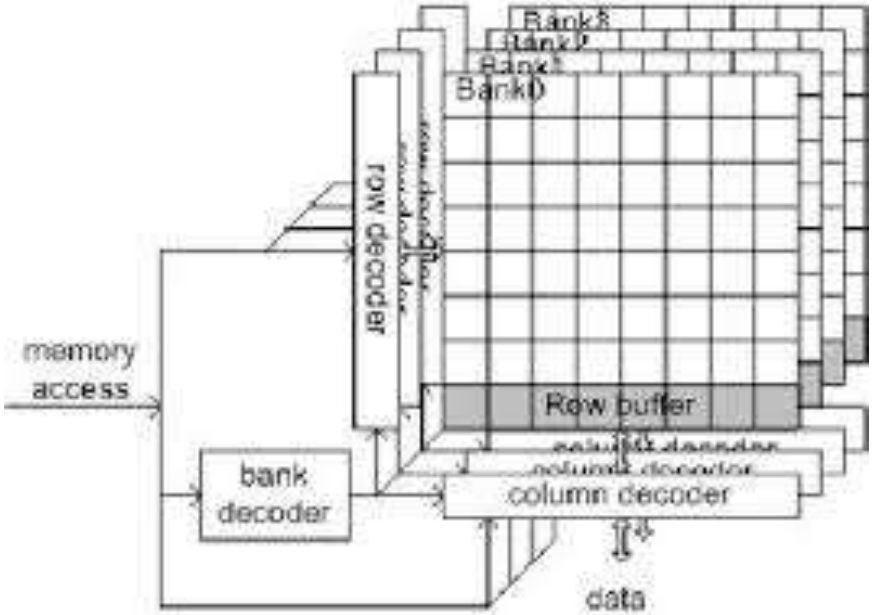
(● = 1, ❍ = 0)



Both types are volatile (ie need power to retain data values)

How is Memory organized and accessed?

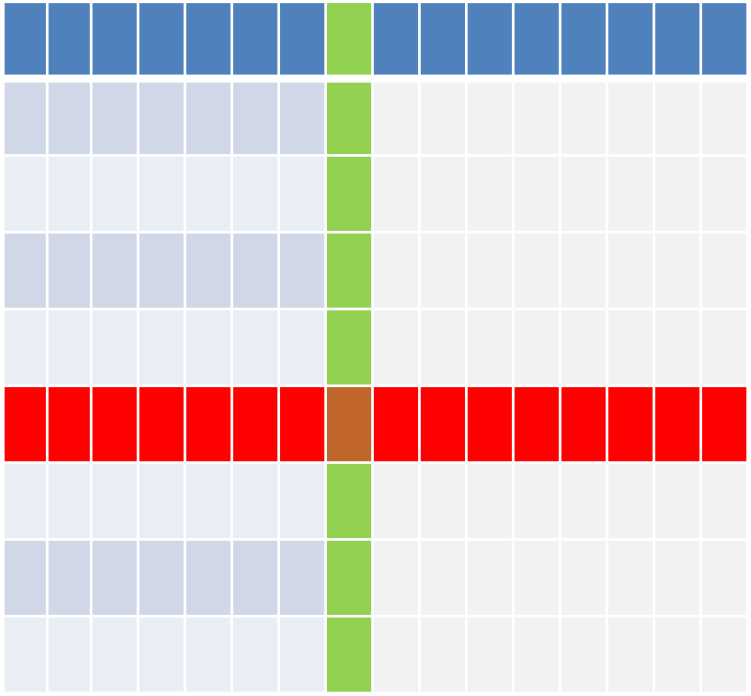
Memory cells are organized into a two dimensional array



Individual cells are specified by a row and column address

Memory Size

- Eg Consider a very simple example using 8 bit addresses (what is the size of the memory?)
- The memory address is 01010111
- 0101 is the row address
- 0111 is the column address



Memory Row Column Access

- Select the Row.

  This causes a read of all data in a given row into a row buffer (Column sense amplifiers). The time for this to take place is called *Row to Column Delay* or t_{RCD} , and is typically specified in number of clock cycles.

- Select the column

  Actually the row operation selects (opens) the entire row, so it is a matter of filtering out the required data (a type of masking operation)

Both of these assumes that the memory bank is ready to be accessed, which is not always the case….

Memory Timing parameters

Memory timings are normally given as 4 numbers,
eg 8-8-8-24
t

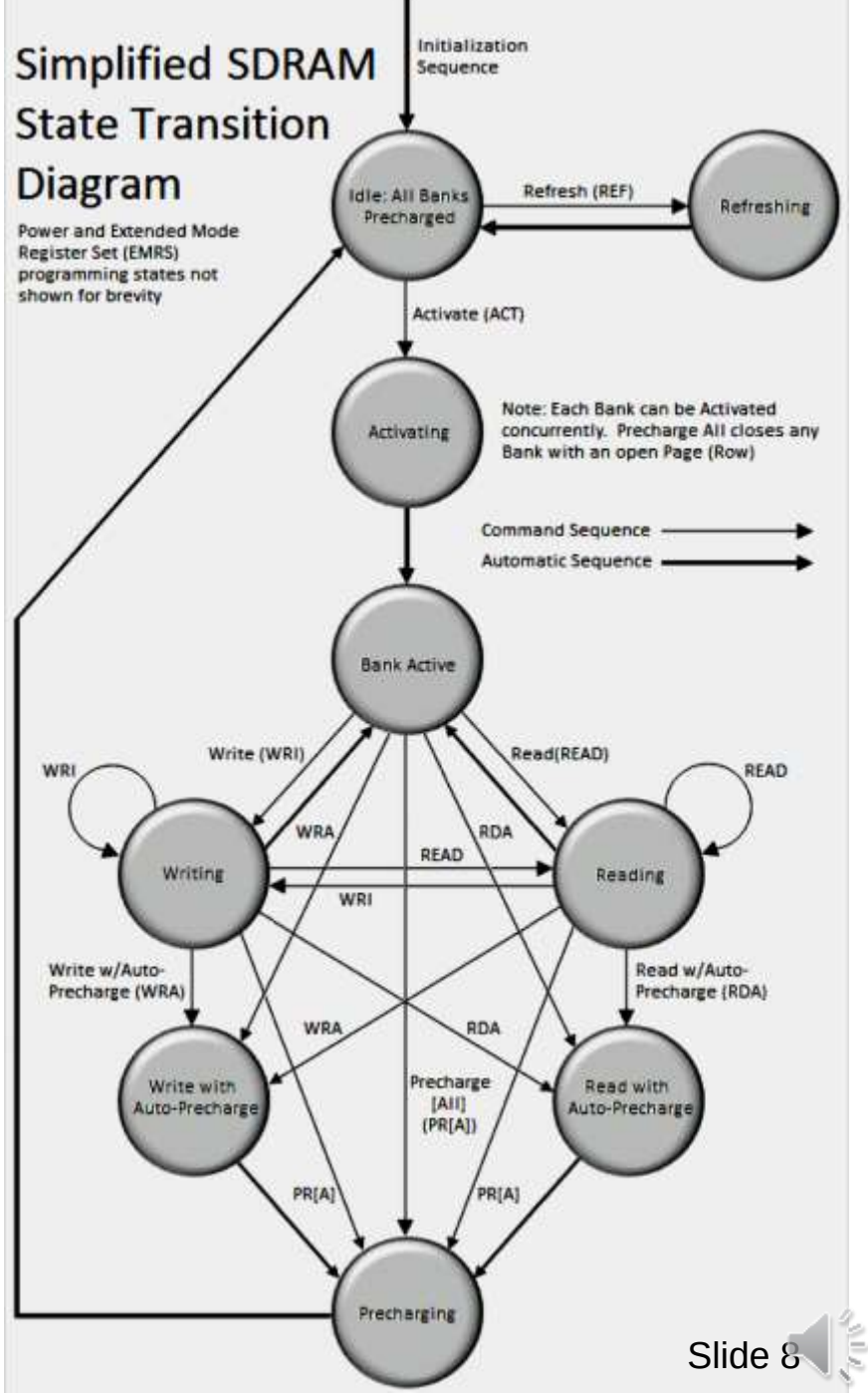
Name	Symbol	Definition
CAS latency	**CL**	The number of cycles between sending a column address to the memory and the beginning of the data in response. This is the number of cycles it takes to read the first bit of memory from a DRAM with the correct row already open. Unlike the other numbers, this is not a maximum, but an exact number that must be agreed on between the memory controller and the memory. **Assumes access to current memory row**
Row Address to Column Address Delay	$\mathbf{T_{RCD}}$	The minimum number of clock cycles required between opening a row of memory and accessing columns within it. The time to read the first bit of memory from a DRAM without an active row is T_{RCD} + CL.
Row Precharge Time	$\mathbf{T_{RP}}$	The minimum number of clock cycles required between issuing the precharge command and opening the next row. The time to read the first bit of memory from a DRAM with the wrong row open is T_{RP} + T_{RCD} + CL. **A row being closed must be precharged**
Row Active Time	$\mathbf{T_{RAS}}$	The minimum number of clock cycles required between a row active command and issuing the precharge command. This is the time needed to internally refresh the row, and overlaps with T_{RCD}. In SDRAM modules, it is simply T_{RCD} + CL. Otherwise, approximately equal to T_{RCD} + $2 \times$ CL.
Notes:		

From Wikipedia

- Row Precharge
  - Despite the name, Row Precharge is done when a row of memory is closed. It essentially regularizes the charges (which represent 0's and 1's) to the correct voltages – These actually change slightly as a result of being read or written
- Row active time
  - A memory row is normally in the idle state and it must be made active (or opened) before it can be read from or written to
- However
  - When a new row is opened, it means that the previous row must be closed, meaning that both pre-charge and row active operations need to be performed
- Some of these operations can be performed in parallel.

Simplified SDRAM State Transition Table

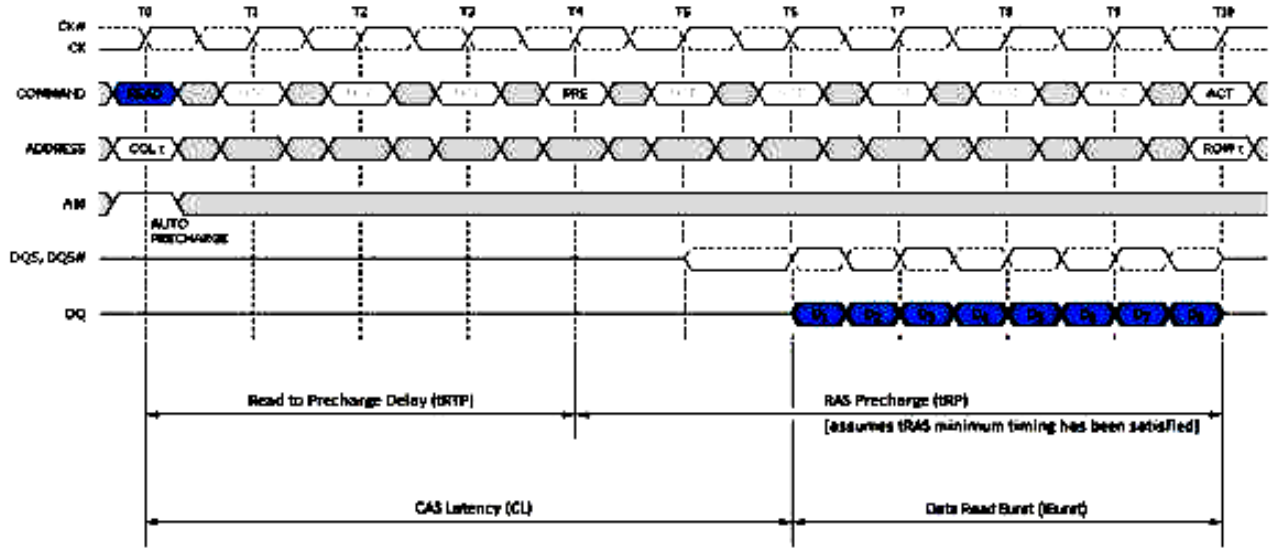
- Using SDRAM is a complex transition between many system states.



Examples of Timing

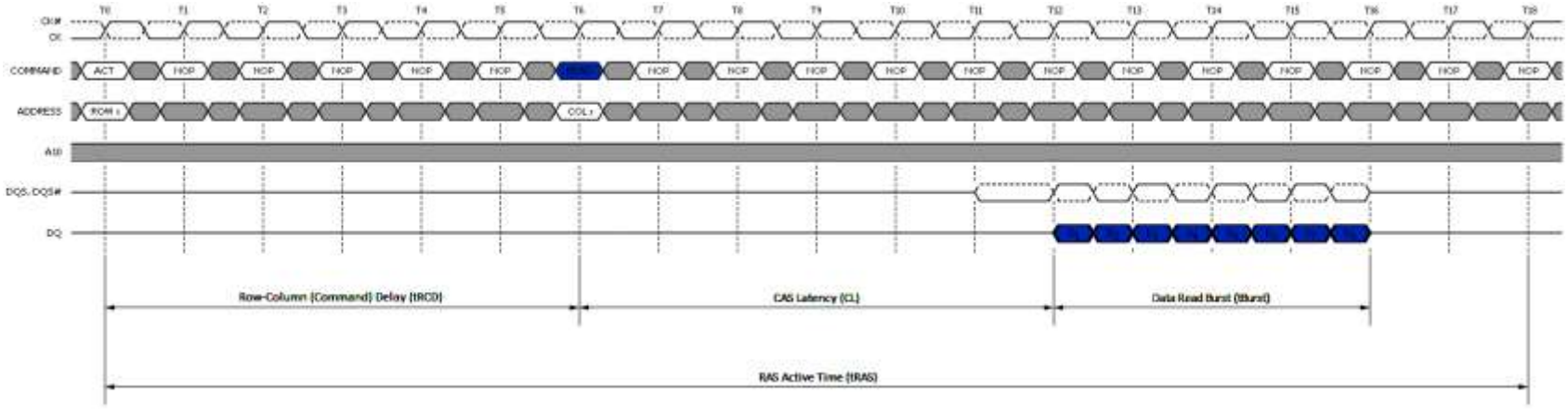
- Assuming 6-8-8-24 timing

- If the required data is in the currently opened memory row, access time is CL, ie 6 clock cycles

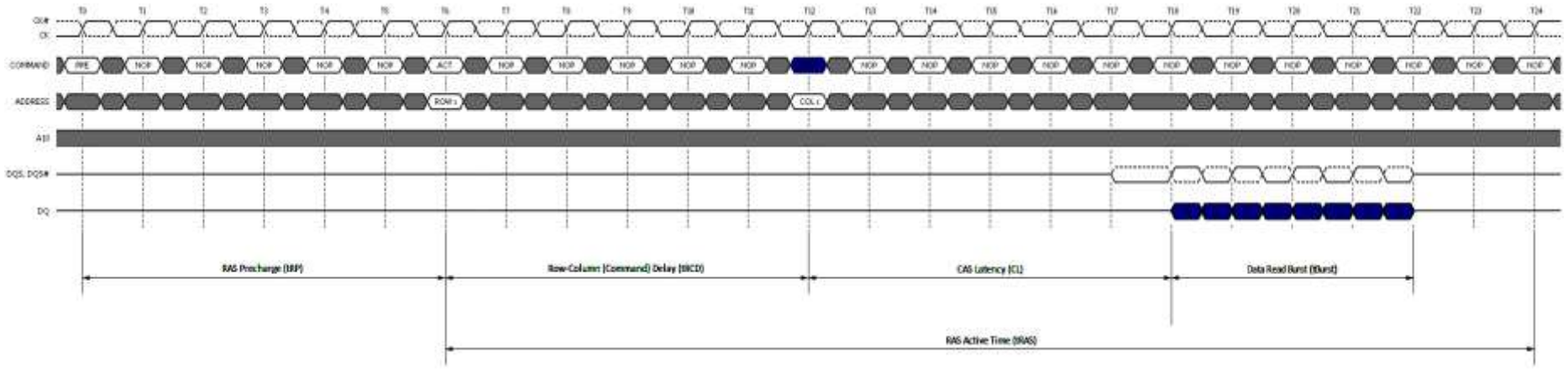


Timing examples from https://www.anandtech.com/show/3851/everything-you-always-wanted-to-know-about-sdram-memory-but-were-afraid-to-ask/5

- If the row needs to be opened first, then the, access time is t_{RCD} + CL, ie 6+6 = 12 clock cycles



- If another row is open, it needs to be closed first, so the total time is $T_{RP} + T_{RCD} + CL = 18$ clock cycles



Banks, Ranks

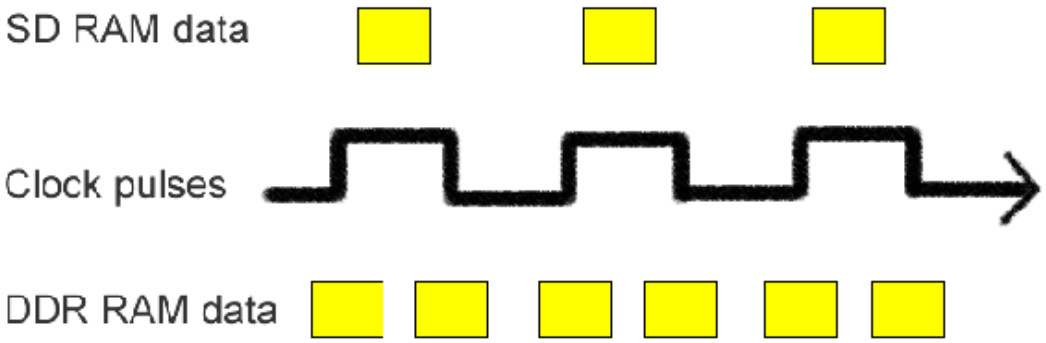
- We have only covered the basics here
- Not covered: memory banks, and memory ranks

Evolution of SDRAM

- Synchronous Dynamic Random-Access Memory (SDRAM)
  - synchronized with system bus
  - pipelined: can start work on new request before previous one finished
- SDRAM - introduced late 1990's
- Ran at 3.3v

Evolution of SDRAM

- DDR SDRAM
- Commonly referred to as DDR ram or DDR1 ram
- The internal clock of the memory modules runs at twice the speed of the CPU clock



DDR1

- Typical DDR SDRAM clock rates are 133, 166 and 200 MHz, which are described as DDR-266, DDR-333 and DDR-400, because of the double clock rate.
- Double clock rate meant that 2 words of data can be read or written every CPU clock cycle.
- Voltage reduced to 2.5v, 184 pin chips
- Peak bandwidth (burst mode) 3.2Gb/sec for DDR-400
  (Memory data bus is 64 bits wide)

DDR2

- Minimum read/write unit doubled to 4 words
- 240 pin chips – mainly additional ground lines
- DDR2 SDRAM clock rates are 200, 266, 333 or 400 MHz, described as DDR2-400, DDR2-533, DDR2-667 and DDR2-800

DDR3

- Minimum read/write unit doubled again to 8 words

- Typical DDR SDRAM clock rates were 400 and 533 MHz, 667MHz and 800 MHz, described as DDR3-800, DDR3-1066, DDR3-1333 and DDR3-1600

- Peak bandwidth (burst mode) 3.2Gb/sec for DDR-400
  (Memory data bus is 64 bits wide)

DDR4

- Designed spec in 2008, released in 2011/12
- Minimum read/write unit 8 words – same as DDR3
- Clock frequencies increased again - 800 to 2133 MHz, described as DDR4-1600 through DDR4-4266

Voltage dropped to1.2-1.4v

- Peak bandwidth (burst mode) 3.2Gb/sec for DDR-400
  (Memory data bus is 64 bits wide)

DDR5

Feature	DDR4	DDR5	DDR5 Advantage
Data rates	3200 Mhz	4800-6400mhz	Increased performance and bandwidth
VDD/VDDQ/VPP	1.2	1.1	Lower power
Device densities	16GB Max	64GB	Larger monolithic devices
Prefetch	8n	16n	Keeps the internal core clock low
CRC	Write	Read/Write	Strengthens system RAS by protecting read data
Burst length	BL8	BL16, BL32	Allows 64B cache line fetch w/only 1 DIMM subchannel.

Current forecast is to start production late 2019 or 2020

- but still no product released

Summary

- Memory
  - Technologies
  - Timing
  - Evolution of SDRAM
- Main Memory
- Memory Cache