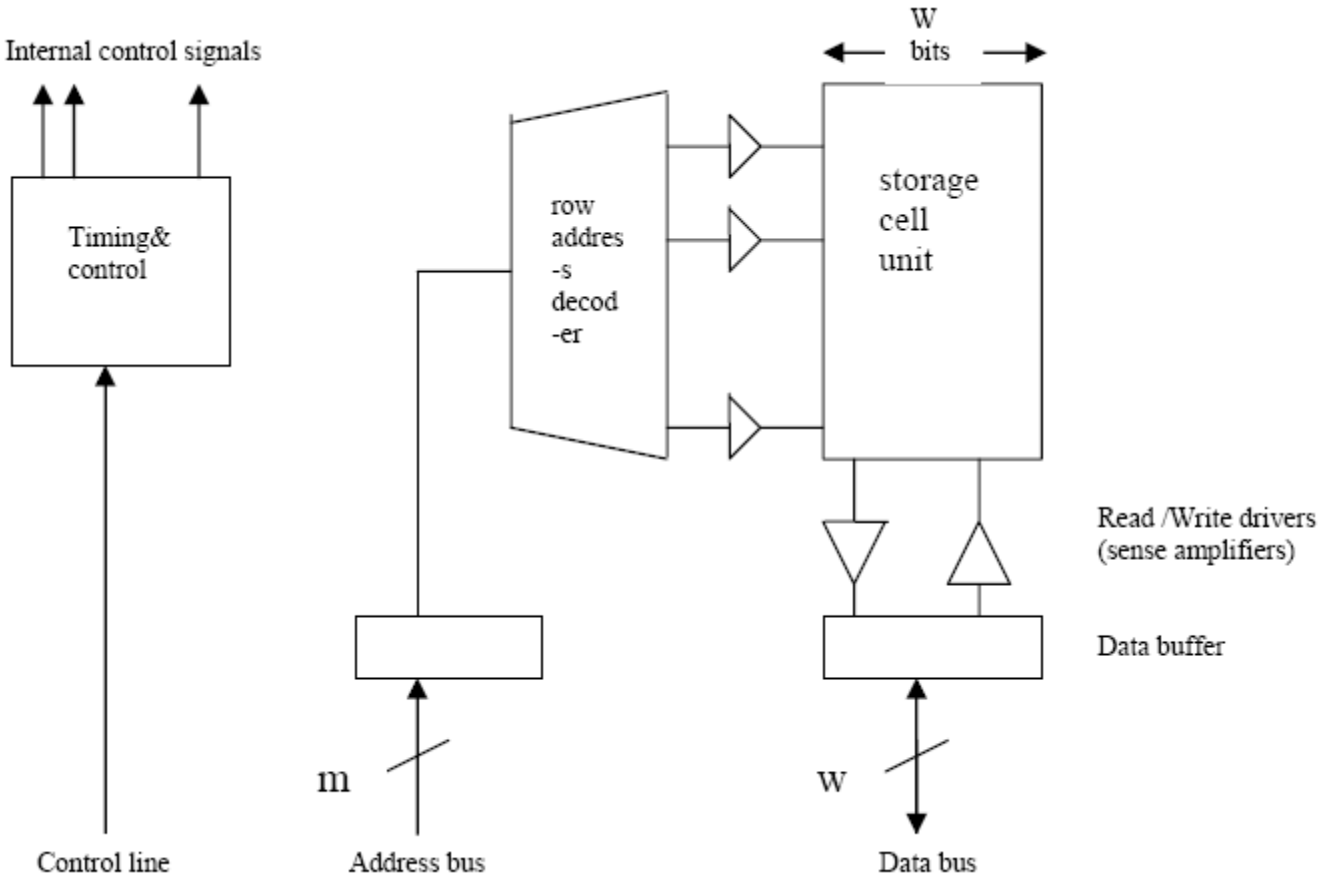
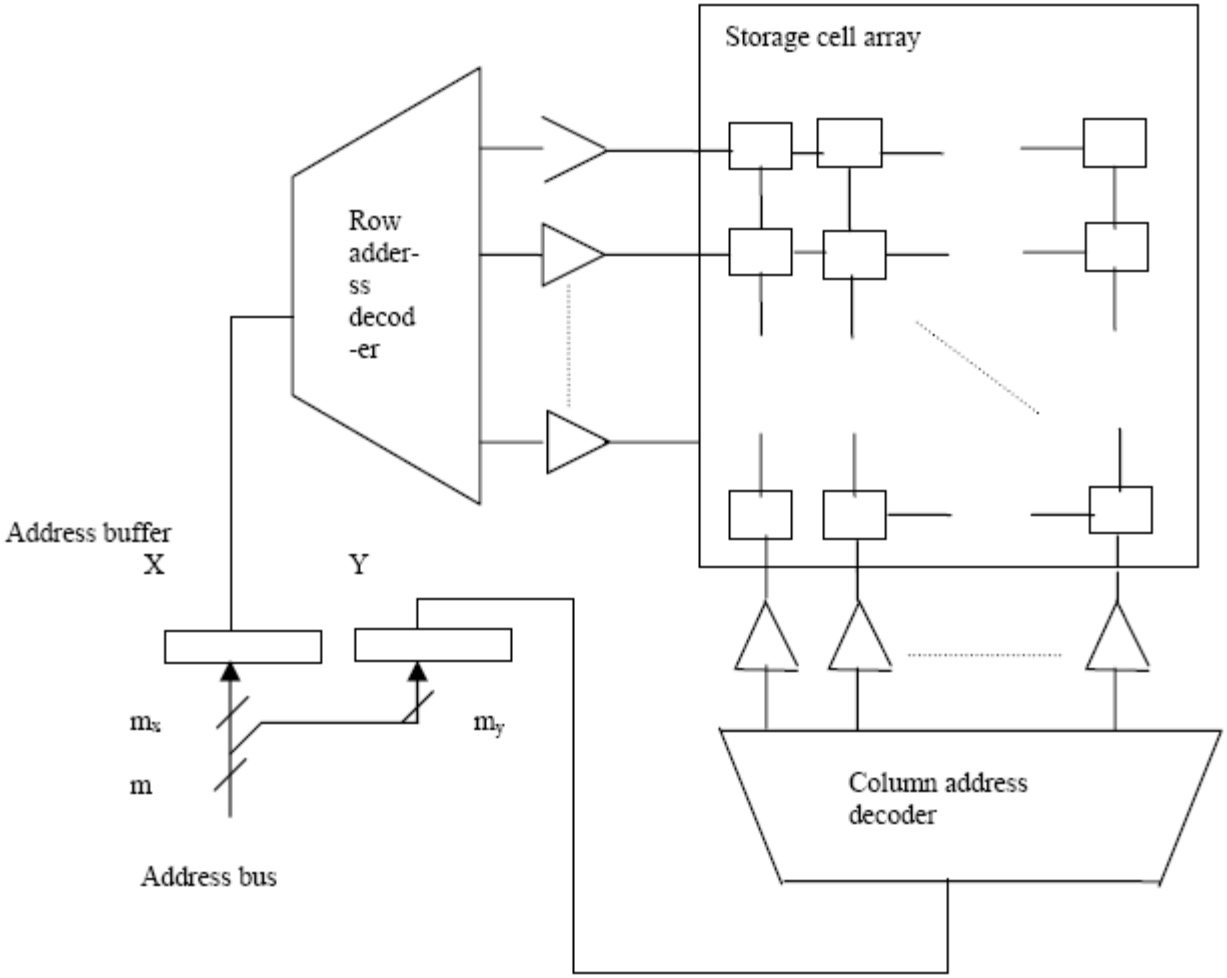
COMPUTER ORGANIZATION (IS F242)

LECT 49: DRAM

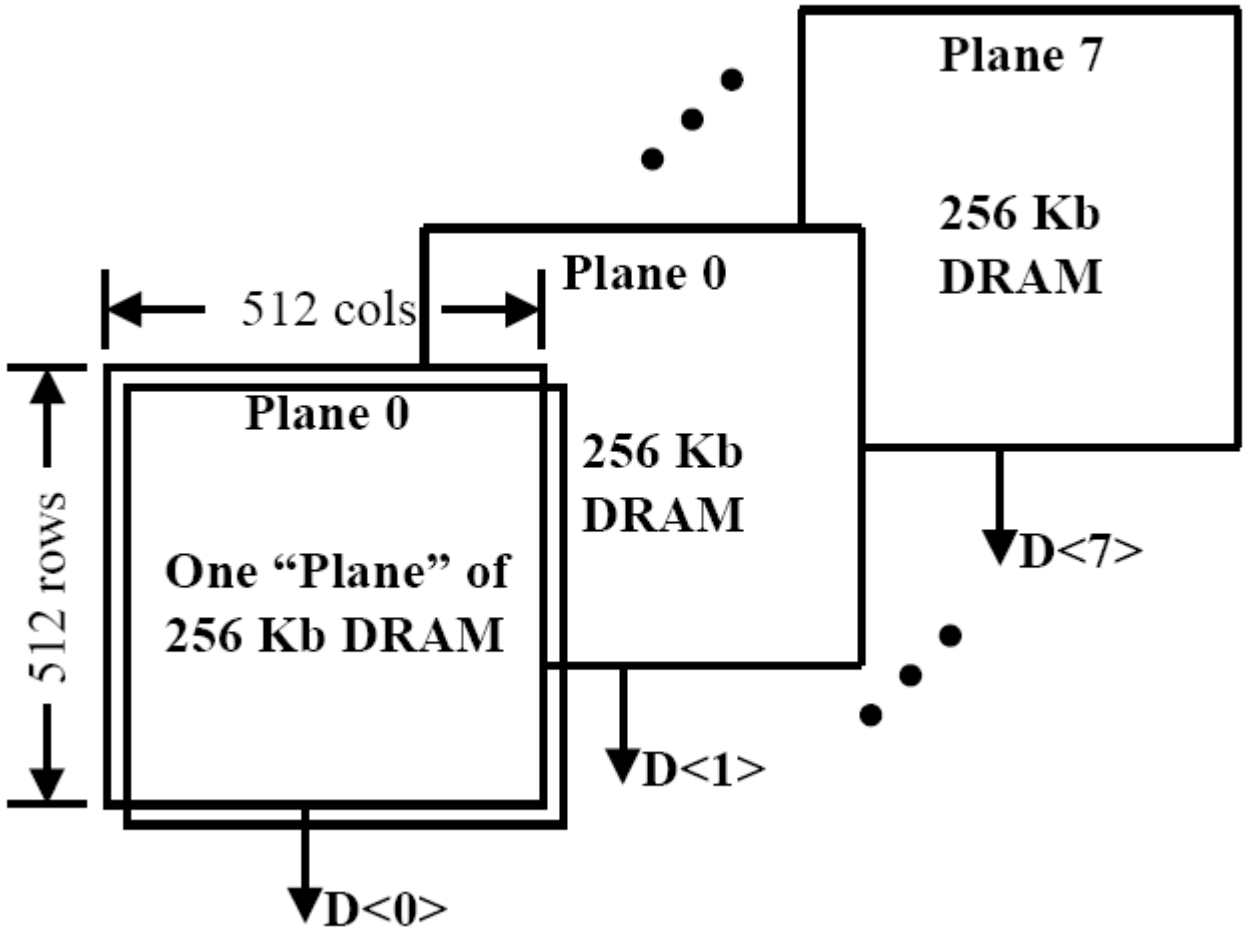
1-D


Internal control signals
Timing&
control
Control line
W
bits
row
addres
-s
decod
-er
storage
cell
unit
Read /Write drivers
(sense amplifiers)
Data buffer
m
Address bus
w
Data bus

2-D


Storage cell array
Row
adder-
ss
decod
-er
Address buffer
X
Y
m_x
m_y
m
Address bus
Column address
decoder

2 Mb DRAM = 256K x 8 = 512 rows x 512 cols x 8 bits



2.5D Organization

- 2.5D organization (used in DRAM chips)
  - Half of the address bits select a row of the square array
  - Whole row of bits is brought out of the memory array into a buffer register (slow, 60-80% of access time)
  - Other half of address bits select one bit of buffer register (with the help of multiplexer), which is read or written
  - Whole row is written back to memory array
  - Organization demanded by needs of refresh
  - Has other advantages such as nibble, page, and static column mode operation

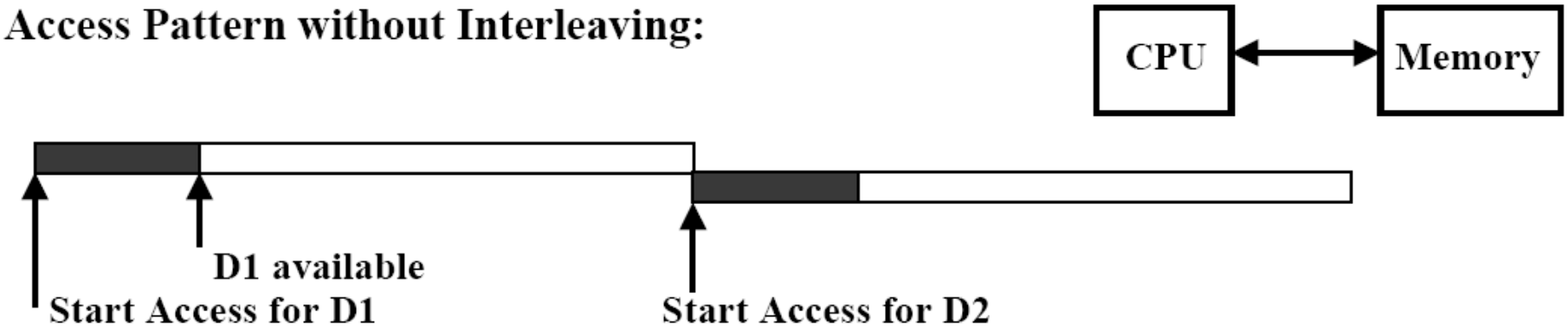
DRAM Refresh

- Consider a 1Mx1 DRAM chip with 190 ns cycle time
- Time for refreshing one bit at a time
  - $190 \times 10^{-9} \times 10^{6} = 190$ ms > 4-8 ms
- Time for refreshing one row at a time
  - $190 \times 10^{-9} \times 10^{3} = 0.19$ ms < 4-8 ms
- Refresh complicates operation of memory
- Refresh control competes with CPU for access to DRAM
- Each row refreshed once every 4-8 ms irrespective of the use of that row
- Want to keep refresh fast (< 5-10% of total time)

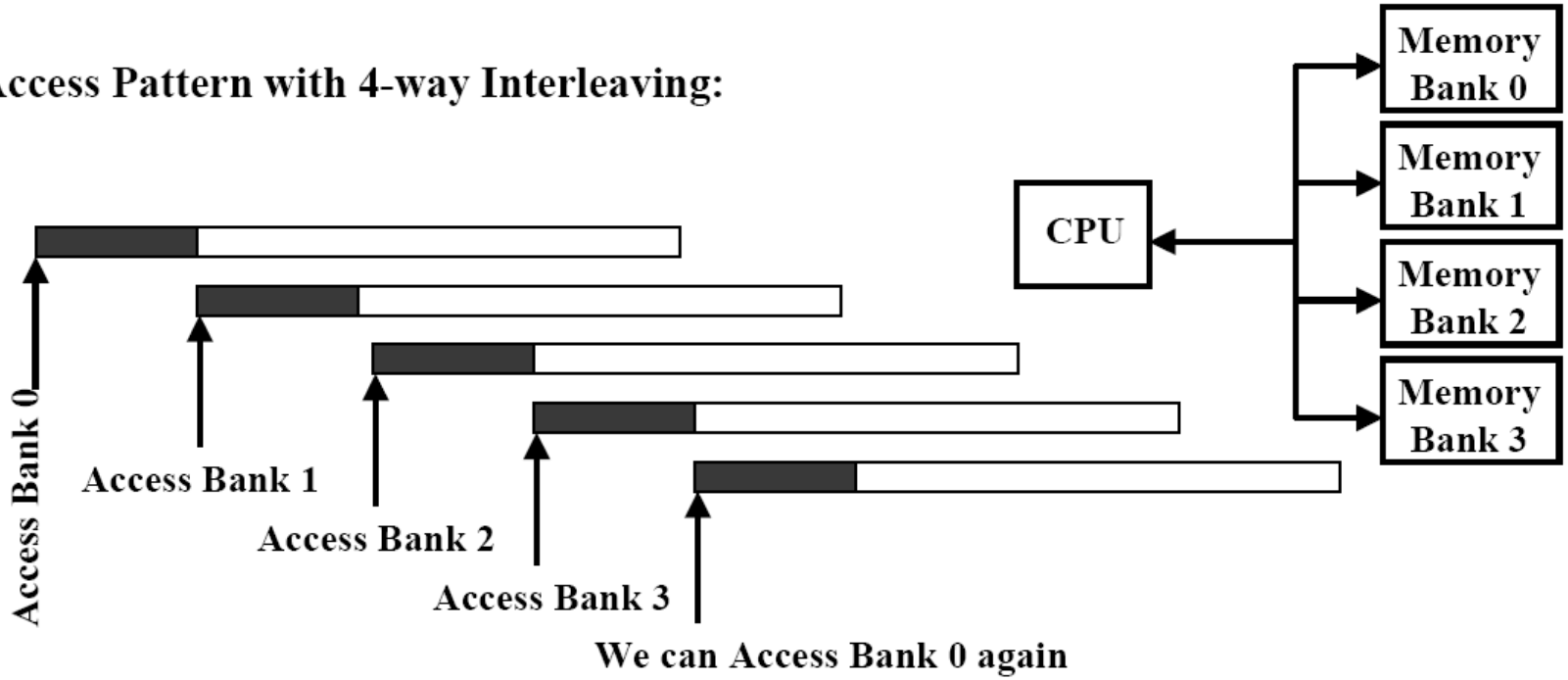
Memory interleaving

- Memory is organized as modules (Banks)
- Tries to take advantage of bandwidth of multiple DRAMs in memory system
- Memory address *A* is converted into (*b*,*w*) pair, where
  - *b* = bank index
  - *w* = word index within bank

Access Pattern without Interleaving:

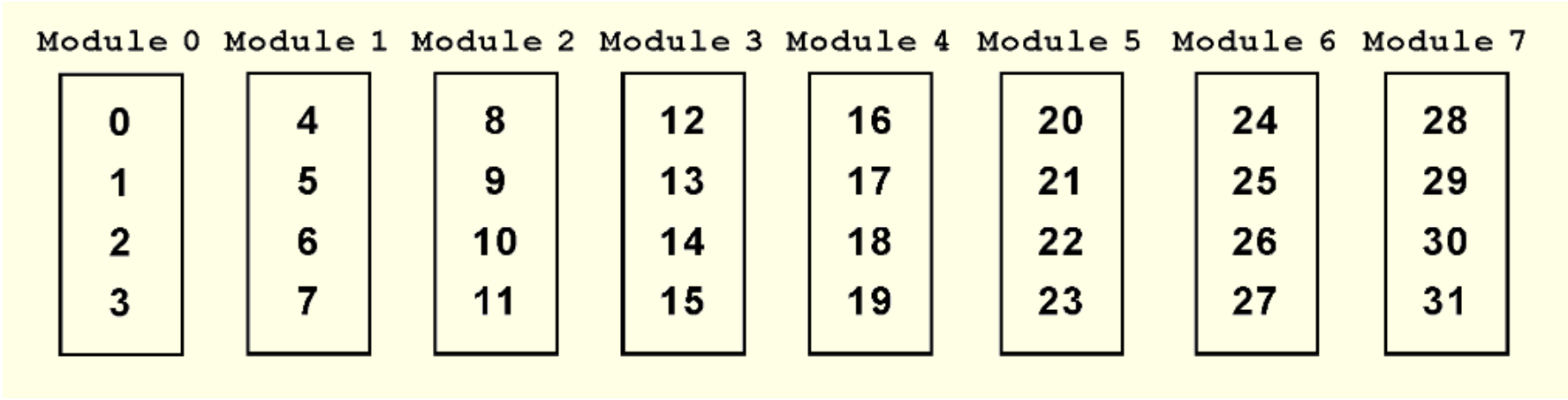


Access Pattern with 4-way Interleaving:



High order interleaving

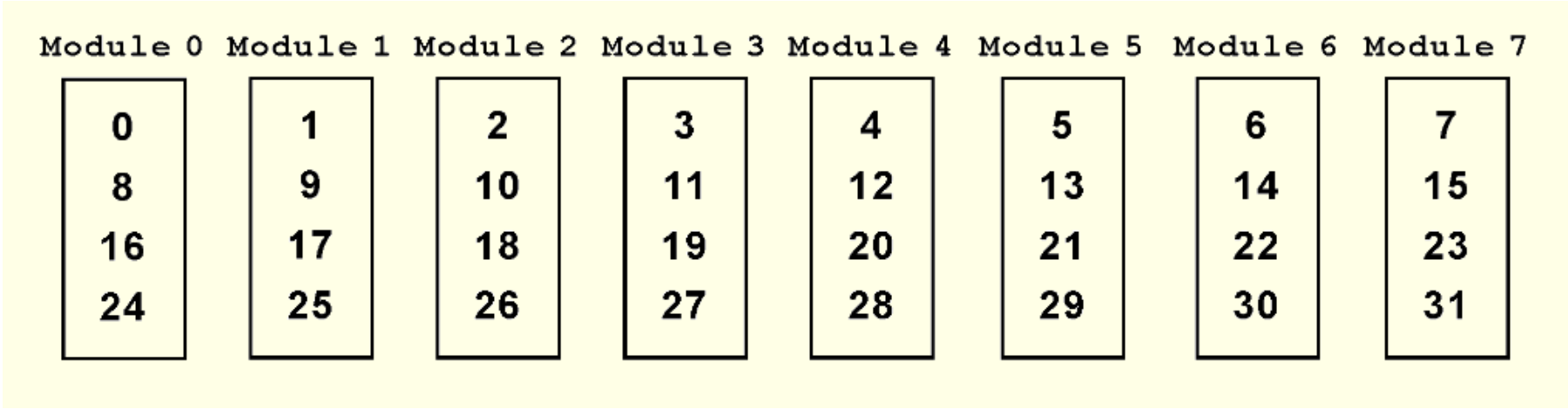
- High order interleaving
  - Uses the high-order bits as the module address
  - Contiguous memory locations assigned to the same bank.



Memory interleaving

- Low order interleaving
  - Uses the low order bits as the module address
  - Access to items from different banks is simultaneous but access to same bank causes a “memory-bank conflict”

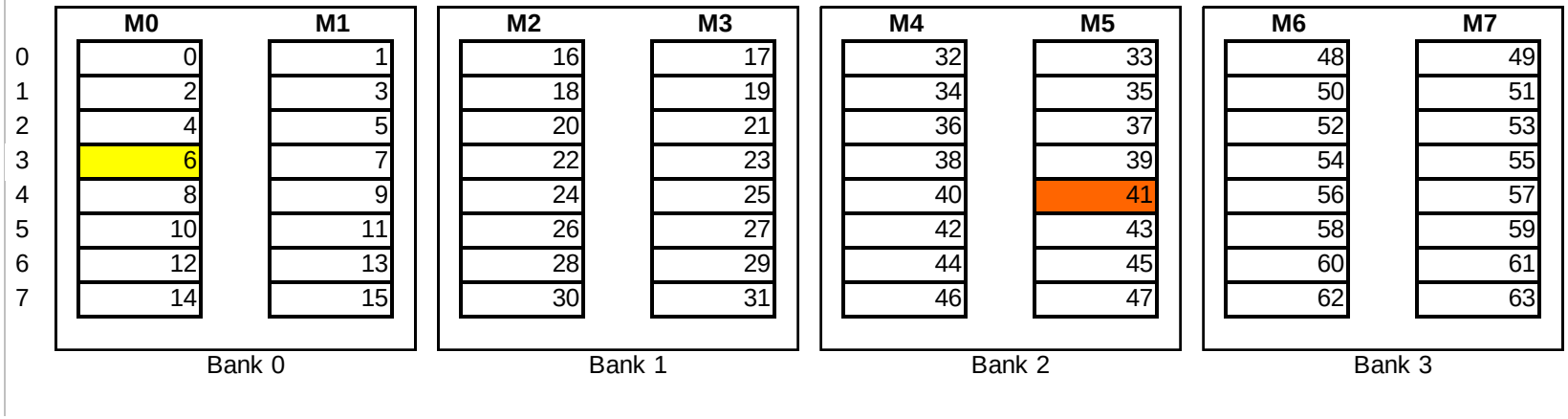
Low order interleaving



Usefulness of Low order interleaving

- If the stride is 1 then Low order interleaving provides maximum parallelism (m parallel memory accesses, where m is the number of banks)
- Useful with the system having cache memory
  - Block replacement has a stride of 1 (quick replacement)

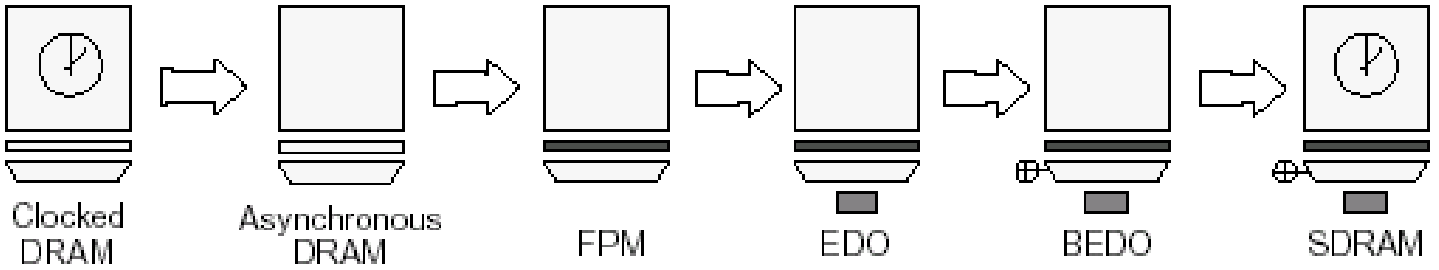
Mixed Interleaving



	M0	M1	M2	M3	M4	M5	M6	M7
0	0	1	16	17	32	33	48	49
1	2	3	18	19	34	35	50	51
2	4	5	20	21	36	37	52	53
3	6	7	22	23	38	39	54	55
4	8	9	24	25	40	41	56	57
5	10	11	26	27	42	43	58	59
6	12	13	28	29	44	45	60	61
7	14	15	30	31	46	47	62	63
	Bank 0		Bank 1		Bank 2		Bank 3	

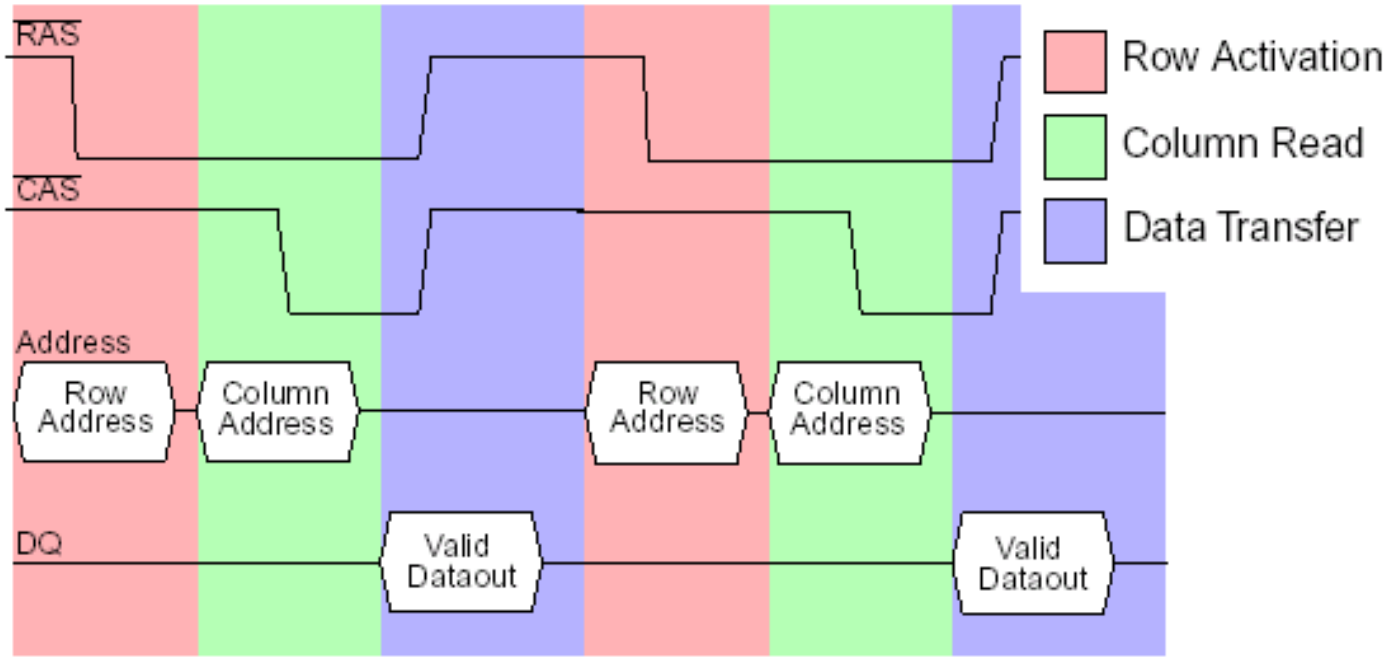
- Memory address register is 6 bits wide
  - Most significant 2 bits give bank address
  - Next 3 bits give word address within bank
  - LSB gives (parity of) module within bank
- $6 = 000110_2 = (00, 011, 0) = (0, 3, 0)$
- $41 = 101001_2 = (10, 100, 1) = (2, 4, 1)$

Evolution of DRAM Architecture



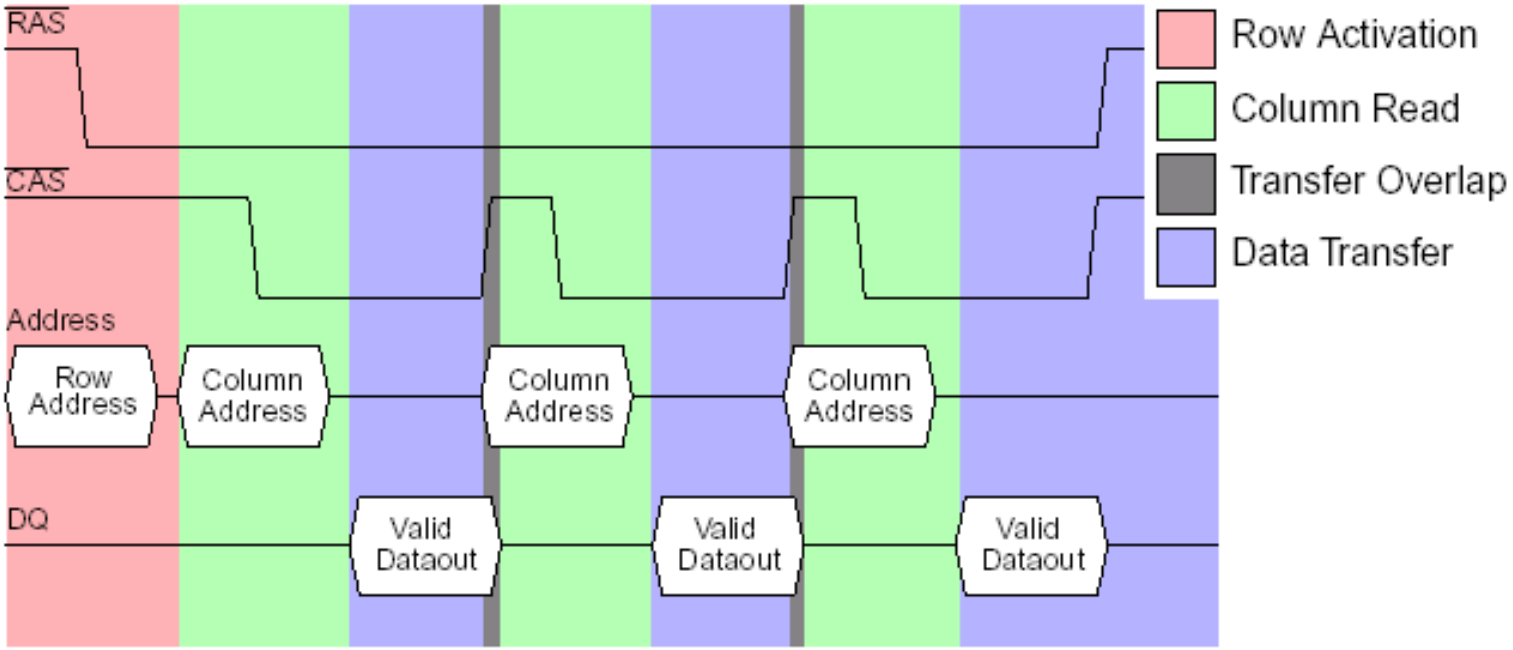
DRAMS

- Conventional DRAM
  - RAS, CAS, RAS, CAS,…
- Read Timings of Conventional DRAMS



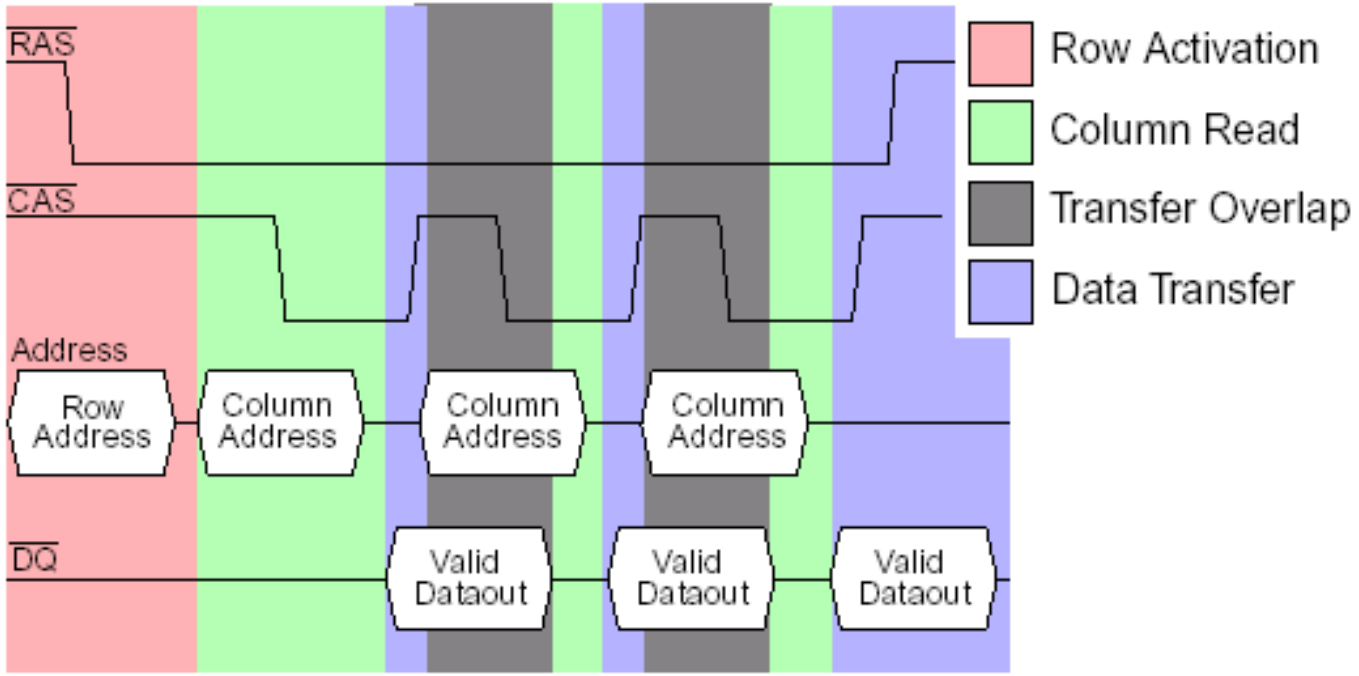
DRAMS

- FPM DRAM
  - Fast Page Mode
  - RAS, CAS, CAS, CAS,…
  - Upto 33MHz
- FPM DRAM Read Timing



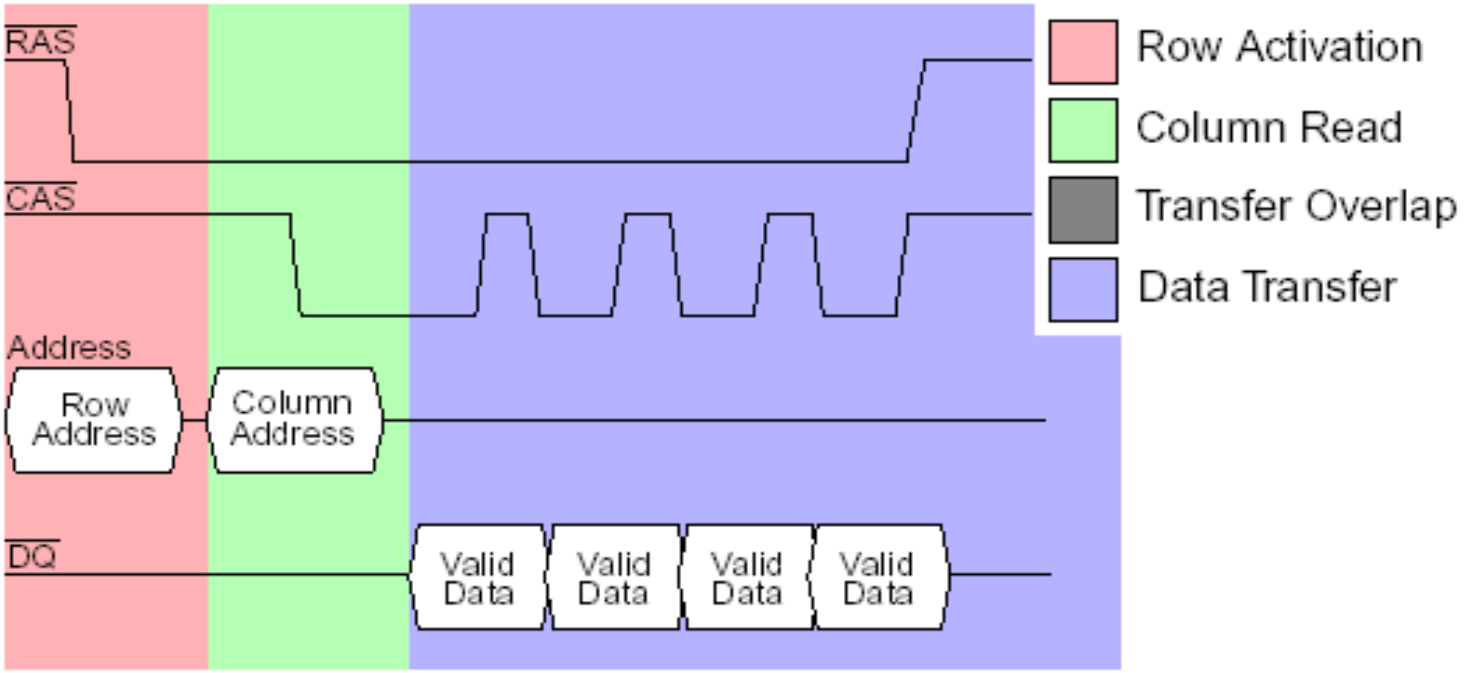
DRAMS

- EDO (Extended Data Out) DRAM
  - Extended Data Out
  - 5-2-2-2
- EDO DRAMS Read Timing



DRAMS

- BEDO DRAM
  - Burst Extended Data Out
  - 5-1-1-1
- Burst EDO Read Timing



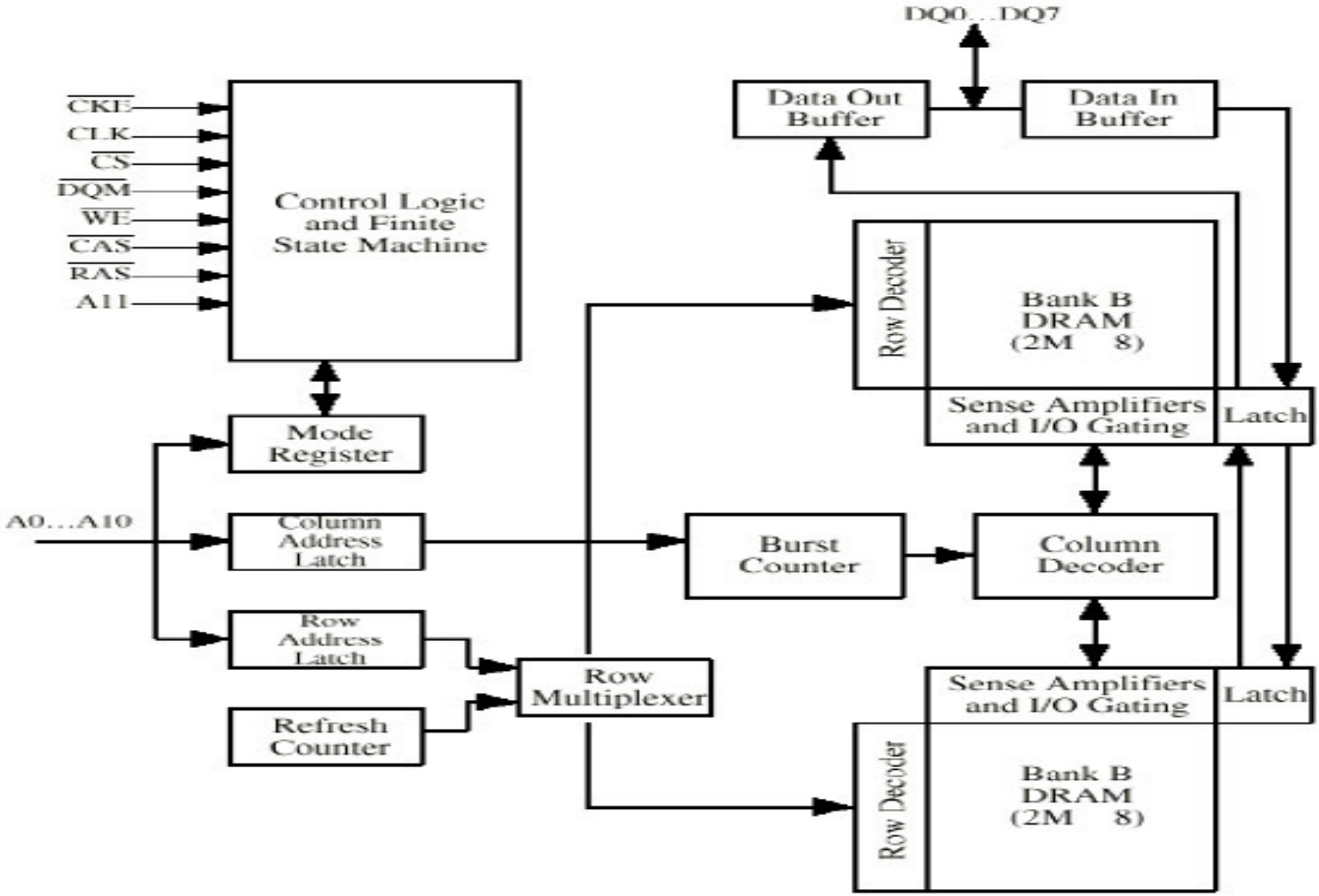
Evolutionary DRAM Architectures

- SDRAM (Synchronous DRAM)
  - All address, data, and control signals are synchronized with an external clock (100-150 MHz)
    - Allows decoupling of processor and memory
    - Allows pipelining a series of reads and writes
  - Peak speed per memory module: 800-1200 MB/sec

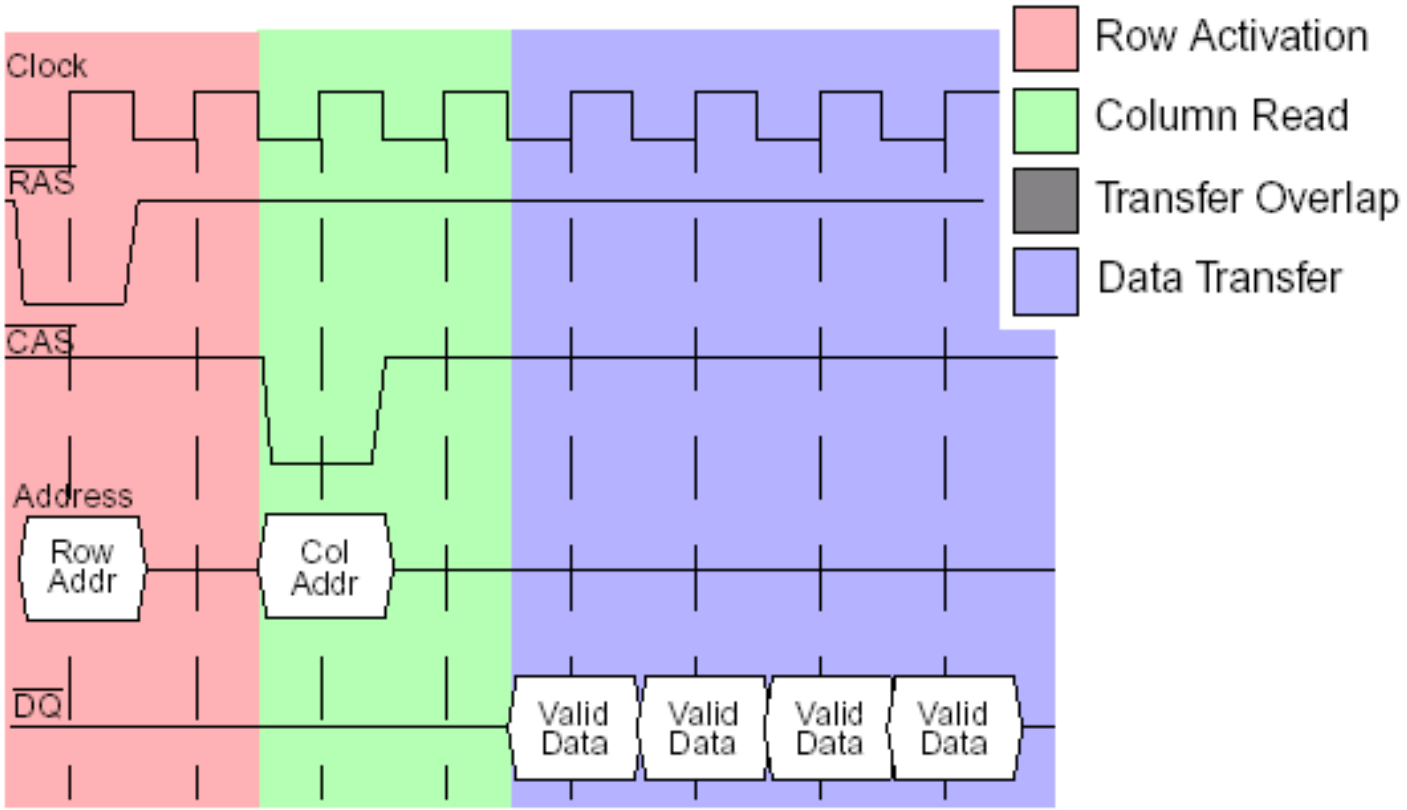
DRAMS

- Synchronous DRAM (SDRAM)
  - Overcome the over head to synchronize with the controller.
  - Access is synchronized with an external clock
  - Address is presented to RAM
  - RAM finds data (CPU waits in conventional DRAM)
  - Since SDRAM moves data in time with system clock, CPU knows when data will be ready
  - CPU does not have to wait, it can do something else
  - Burst mode allows SDRAM to set up stream of data and fire it out in block
  - 4-1-1-1

SDRAM


DQ0...DQ7
Data Out Buffer
Data In Buffer
CKE
CLK
CS
DQM
WE
CAS
RAS
A11
Control Logic and Finite State Machine
Row Decoder
Bank B DRAM (2M 8)
Sense Amplifiers and I/O Gating
Latch
Mode Register
A0...A10
Column Address Latch
Burst Counter
Column Decoder
Row Address Latch
Row Multiplexer
Refresh Counter
Sense Amplifiers and I/O Gating
Latch
Row Decoder
Bank B DRAM (2M 8)

SDR SDRAM Read Timing



DRAMS

- SDRAM(continue)
  - Can work at 100MHz / 133 MHz
  - PC100 / PC 133
  - Memory Interleaved, Dual Bank
  - On-Chip Parallelism
- DDR SDRAM
  - Dual Data Rate SDRAM
  - Access at both clock edges
  - Dual-Bank Technology, same as SDRAMs

- DDR2 SDRAM
  - Higher bus speed
  - Doubles the bus frequency
  - Not compatible with DDR SDRAM
  - The best-rated DDR2 memory modules are at least twice as fast as the best-rated DDR memory modules.

Production year	Chip size	DRAM Type	Row access strobe (RAS)		Column access strobe (CAS)/ data transfer time (ns)	Cycle time (ns)
			Slowest DRAM (ns)	Fastest DRAM (ns)		
1980	64K bit	DRAM	180	150	75	250
1983	256K bit	DRAM	150	120	50	220
1986	1M bit	DRAM	120	100	25	190
1989	4M bit	DRAM	100	80	20	165
1992	16M bit	DRAM	80	60	15	120
1996	64M bit	SDRAM	70	50	12	110
1998	128M bit	SDRAM	70	50	10	100
2000	256M bit	DDR1	65	45	7	90
2002	512M bit	DDR1	60	40	5	80
2004	1G bit	DDR2	55	35	5	70
2006	2G bit	DDR2	50	30	2.5	60
2010	4G bit	DDR3	36	28	1	37
2012	8G bit	DDR3	30	24	0.5	31

Figure 2.13 Times of fast and slow DRAMs vary with each generation. (Cycle time is defined on page 95.) Performance improvement of row access time is about 5% per year. The improvement by a factor of 2 in column access in 1986 accompanied the switch from NMOS DRAMs to CMOS DRAMs. The introduction of various burst transfer modes in the mid-1990s and SDRAMs in the late 1990s has significantly complicated the calculation of access time for blocks of data; we discuss this later in this section when we talk about SDRAM access time and power. The DDR4 designs are due for introduction in mid- to late 2012. We discuss these various forms of DRAMs in the next few pages.

Standard name	Memory clock	Cycle time	I/O Bus clock	Data transfer
DDR2-400	100 MHz	10 ns	200 MHz	400 Million
DDR2-533	133 MHz	7.5 ns	266 MHz	533 Million
DDR2-667	166 MHz	6 ns	333 MHz	667 Million
DDR2-800	200 MHz	5 ns	400 MHz	800 Million
DDR2-1066	266 MHz	3.75 ns	533 MHz	1066 Million

- DDR3
  - DDR3 works by doubling the prefetch data.
  - This is achieved by doubling the internal bank to 8.
  - Each internal bank can process 1 bit of data.
  - Additional internal banks help DDR3 SDRAM to double the data transfer rate without having to make modifications to the original design.
  - The required voltage for DDR3 is 1.5V, a 16% reduction of the DDR2 1.8V (2.5V in case of DDR).
    - 30% reduction in power consumption compared with DDR2 RAM operating at the same speed.
  - Data transfer rate
    - 200–400 (DDR), 400–800 (DDR2) and 800–1600 (DDR3)

Standard name	Memory clock	Cycle time	I/O Bus clock	Data transfer
DDR3-800	100 MHz	10 ns	400 MHz	800 MT/s
DDR3-1066	133 MHz	7.5 ns	533 MHz	1066 MT/s
DDR3-1333	166 MHz	6 ns	667 MHz	1333 MT/s
DDR3-1600	200 MHz	5 ns	800 MHz	1600 MT/s

DDR4

- First DDR4 manufactured by Samsung in 2011
- Expected commercial release by 2014
- Not directly compatible with other DRAMS
  - Different signaling voltages, timings and physical interface
- Higher range of clock frequencies (2133 – 4266 MHz Vs 400 – 800 MHz)
- Higher data transfer rates (3200 Vs 1600)
- Significantly low voltage (1.05 – 1.2 Vs 1.2 – 1.65)
- Uses point to point topology

Thank you for making this a wonderful experience