CS510 Computer Architecture

Lecture 18: DRAMs

Soontae Kim

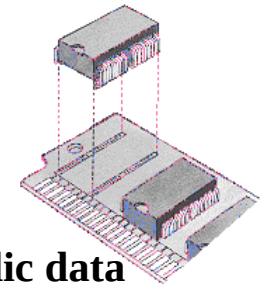
Spring 2017

School of Computing, KAIST

Announcements

- **Homework assignment #3**
  - Posted on class web site
  - Due on May 26 (Friday)
- **Term project**
  - Final presentations on June 7 and 9
  - Final report due on June 16

Main Memory



- **Main memory generally utilizes Dynamic RAM (DRAM),**
  which uses a single transistor and a capacitor to store a bit, but requires periodic data refreshes by reading every row. capacitor discharge over time
- **Static RAM may be used for main memory if the added expense, low density, high power consumption, and complexity is feasible (e.g. Cray Vector Supercomputers).**
- **Main memory performance is affected by:**
  - **Memory latency: Affects cache miss penalty, M. Measured by:**
    - **Access time: The time it takes from the time a memory access request is issued to main memory to the time the requested word is available to cache/CPU.**
    - **Cycle time: The minimum time between requests to memory**

    (greater than access time to allow address lines to be stable)
  - **Peak Memory bandwidth: The maximum sustained data transfer rate between main memory and cache/CPU.**
    - Realistic memory bandwidth < peak memory bandwidth

Logical DRAM Chip Organization

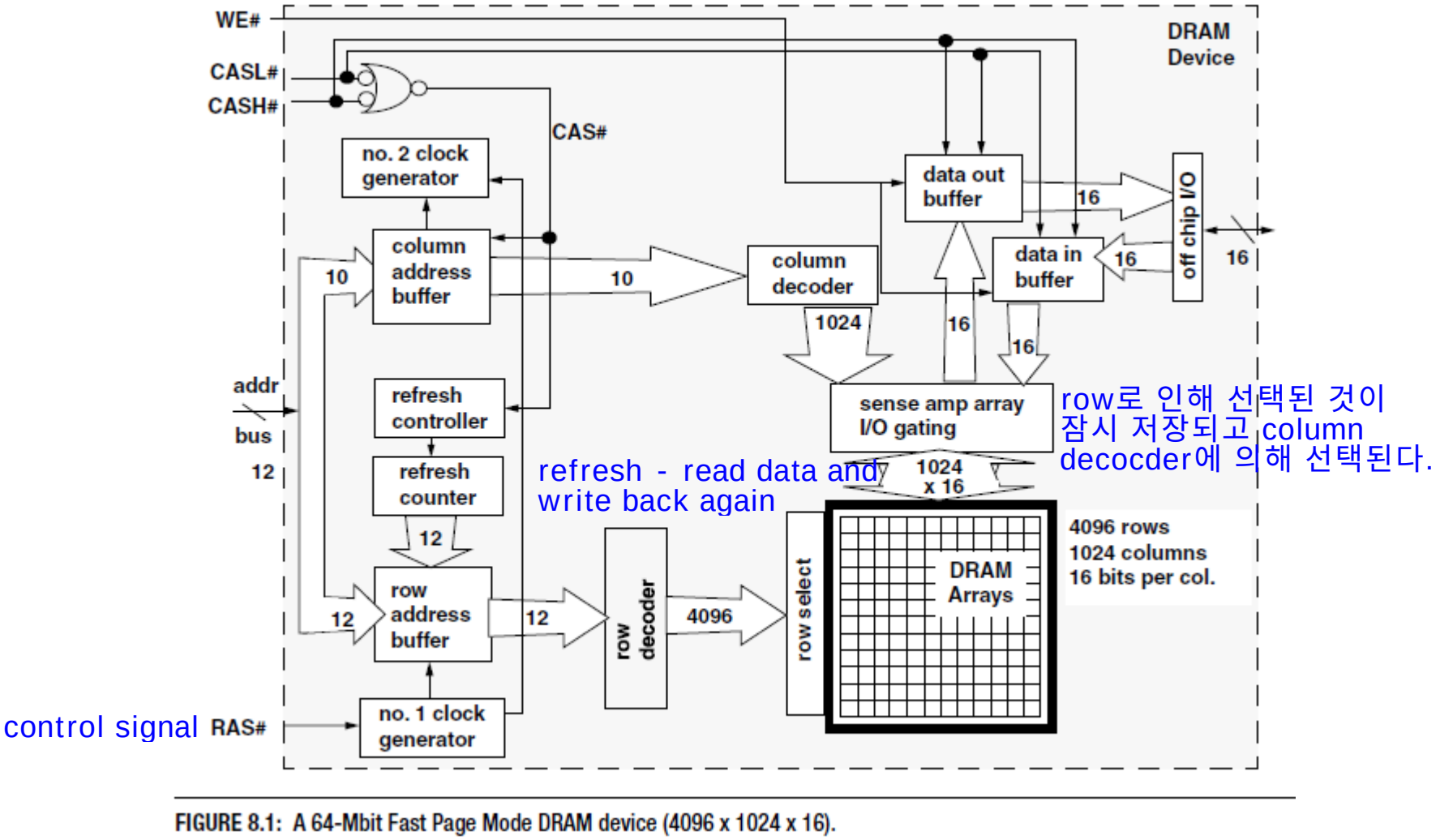


FIGURE 8.1: A 64-Mbit Fast Page Mode DRAM device (4096 x 1024 x 16).

Control Signals:
1 - Row Access Strobe (RAS): Low to latch row address
2- Column Access Strobe (CAS): Low to latch column address

A periodic data refresh is required by reading every bit

Four Key DRAM Timing Parameters

- **t_{RAC}: Minimum time from RAS (Row Access Strobe) line falling (activated) to the valid data output.**
  - **Used to be quoted as the nominal speed of a DRAM chip**
  - **For a typical 64Mb DRAM t_{RAC} = 60 ns**
- **t_{RC}: Minimum time from the start of one row access to the start of the next (memory cycle time).**
  - **t_{RC} = t_{RAC} + RAS Precharge Time** (bitline precharging time & bus reset 등과 같은 electronic관련 issue)
  - **t_{RC} = 110 ns for a 64Mbit DRAM with a t_{RAC} of 60 ns**
- **t_{CAC}: Minimum time from CAS (Column Access Strobe) line falling to valid data output.**
  - **12 ns for a 64Mbit DRAM with a t_{RAC} of 60 ns**
- **t_{PC}: Minimum time from the start of one column access to the start of the next.**
  - **t_{PC} = t_{CAC} + CAS Precharge Time**
  - **About 25 ns for a 64Mbit DRAM with a t_{RAC} of 60 ns**

Simplified DRAM Speed Parameters

- **Row Access Strobe (RAS)Time: (similar to t_{RAC}):**
  - **Minimum time from RAS (Row Access Strobe) line falling (activated) to the first valid data output.**
  - **A major component of memory latency.**
  - **Only improves ~ 5% every year.**
- **Column Access Strobe (CAS) Time/data transfer time: (similar to t_{CAC})**
  - **The minimum time required to read additional data by changing column address while keeping the same row address.**
  - **Along with memory bus width, determines peak memory bandwidth.**

DRAM Generations

Year	Size	RAS (ns)	CAS (ns)	Cycle Time	Memory Type	
1980	64 Kb	150-180	75	250 ns	Page Mode	Asynchronous DRAM
1983	256 Kb	120-150	50	220 ns	Page Mode	
1986	1 Mb	100-120	25	190 ns		
1989	4 Mb	80-100	20	165 ns	Fast Page Mode	
1992	16 Mb	60-80	15	120 ns	EDO	
1996	64 Mb	50-70	12	110 ns	PC66 SDRAM	Synchronous DRAM
1998	128 Mb	50-70	10	100 ns	PC100 SDRAM	
2000	256 Mb	45-65	7	90 ns	PC133 SDRAM	
2002	512 Mb	40-60	5	80 ns	PC2700 DDR	
2004	1Gb	35-55	5	70 ns	PC4300 DDR2	
2006	2Gb	30-50	2.5	60 ns	PC8500 DDR3	
	32000:1 (Capacity)		30:1 (~bandwidth)	4:1 (Latency)	frequency bandwidth	

Basic Memory Bandwidth Improvement/Miss Penalty (M) Reduction Techniques

- **Wider Main Memory (CPU-Memory Bus):**

Memory bus width is increased to a number of words (usually up to the size of a cache block).

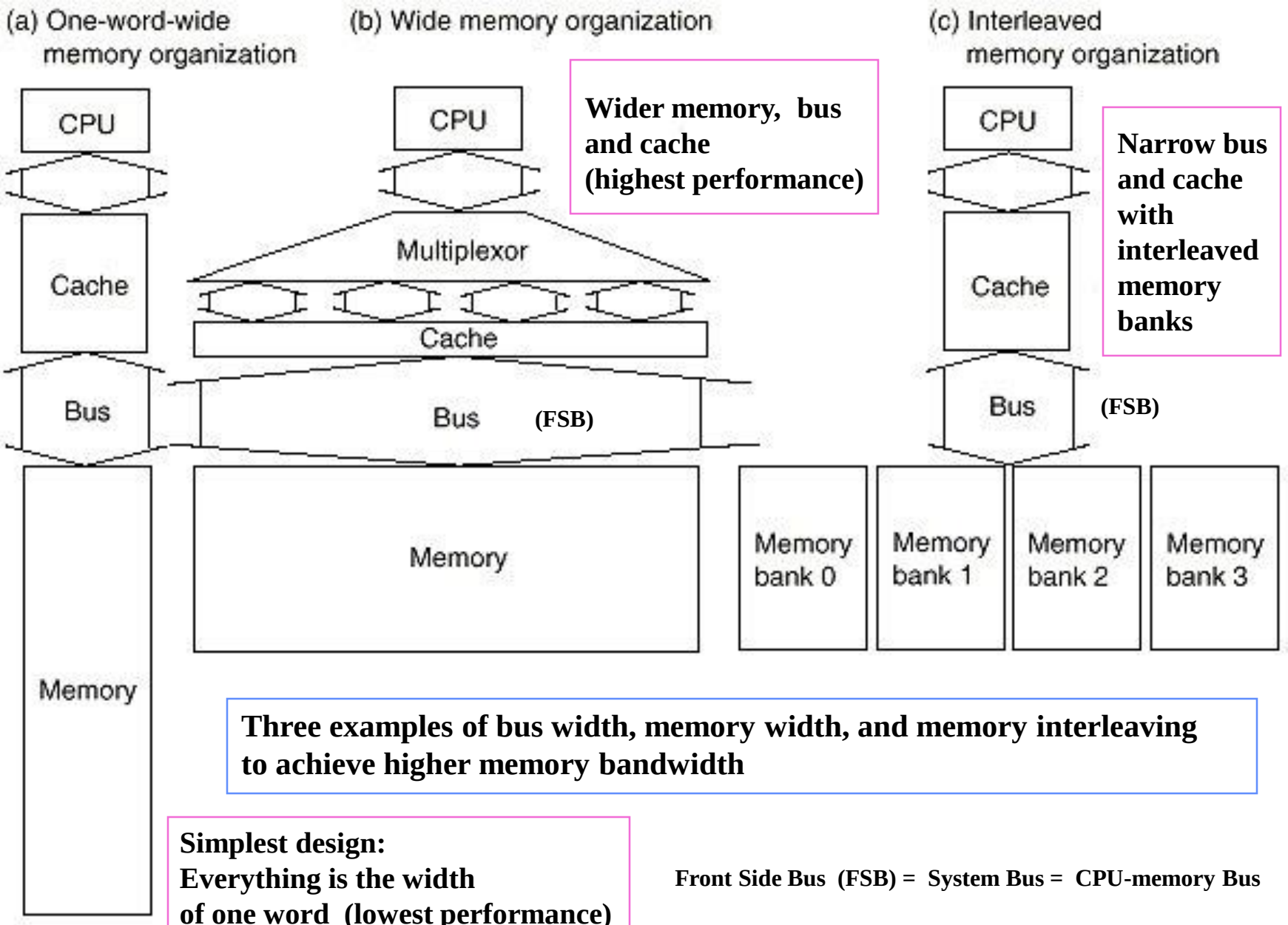
- **Memory bandwidth is proportional to memory bus width.**
  - **e.g Doubling the width of cache and memory doubles potential memory bandwidth available to the CPU.**
- **The miss penalty is reduced since fewer memory bus accesses are needed to fill a cache block on a miss.**

- **Interleaved (Multi-Bank) Memory:**

Memory is organized as a number of independent banks.

- **Multiple interleaved memory reads or writes are accomplished by sending memory addresses to several memory banks at once or pipeline accesses to the banks.**
- **Interleaving factor: Refers to the mapping of memory addressees to memory banks. Goal reduce bank conflicts.**

  e.g. using 4 banks (one word wide), bank 0 has all words whose addresses are: (word address mod) 4 = 0



Wider memory, bus and cache (highest performance)

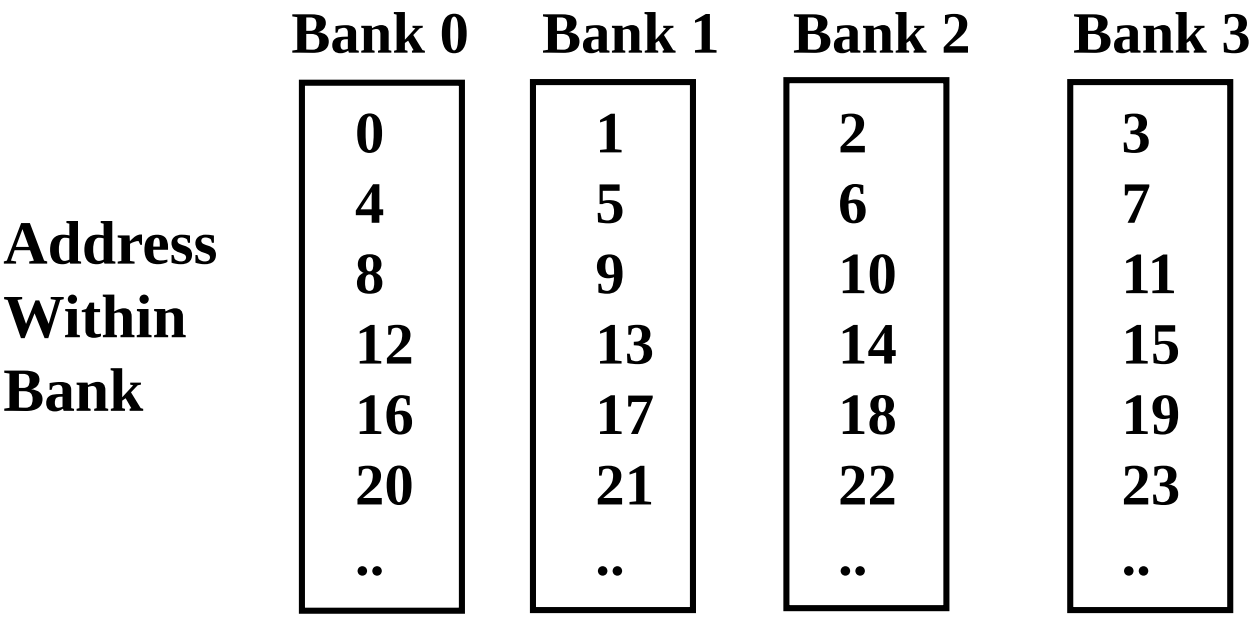
Narrow bus and cache with interleaved memory banks

Three examples of bus width, memory width, and memory interleaving to achieve higher memory bandwidth

Simplest design:
Everything is the width of one word (lowest performance)

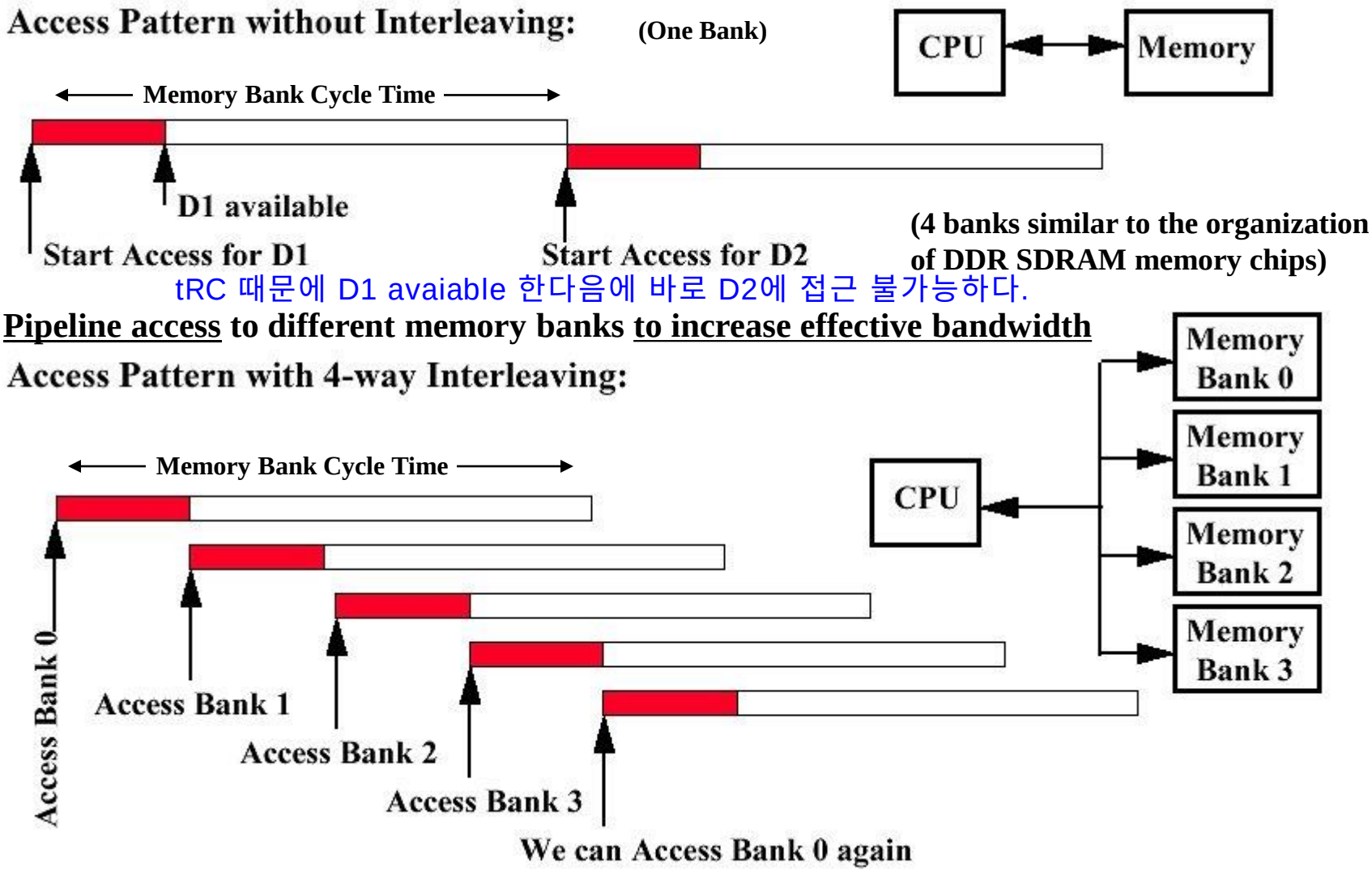
Front Side Bus (FSB) = System Bus = CPU-memory Bus

Four Way (Four Banks) Interleaved Memory



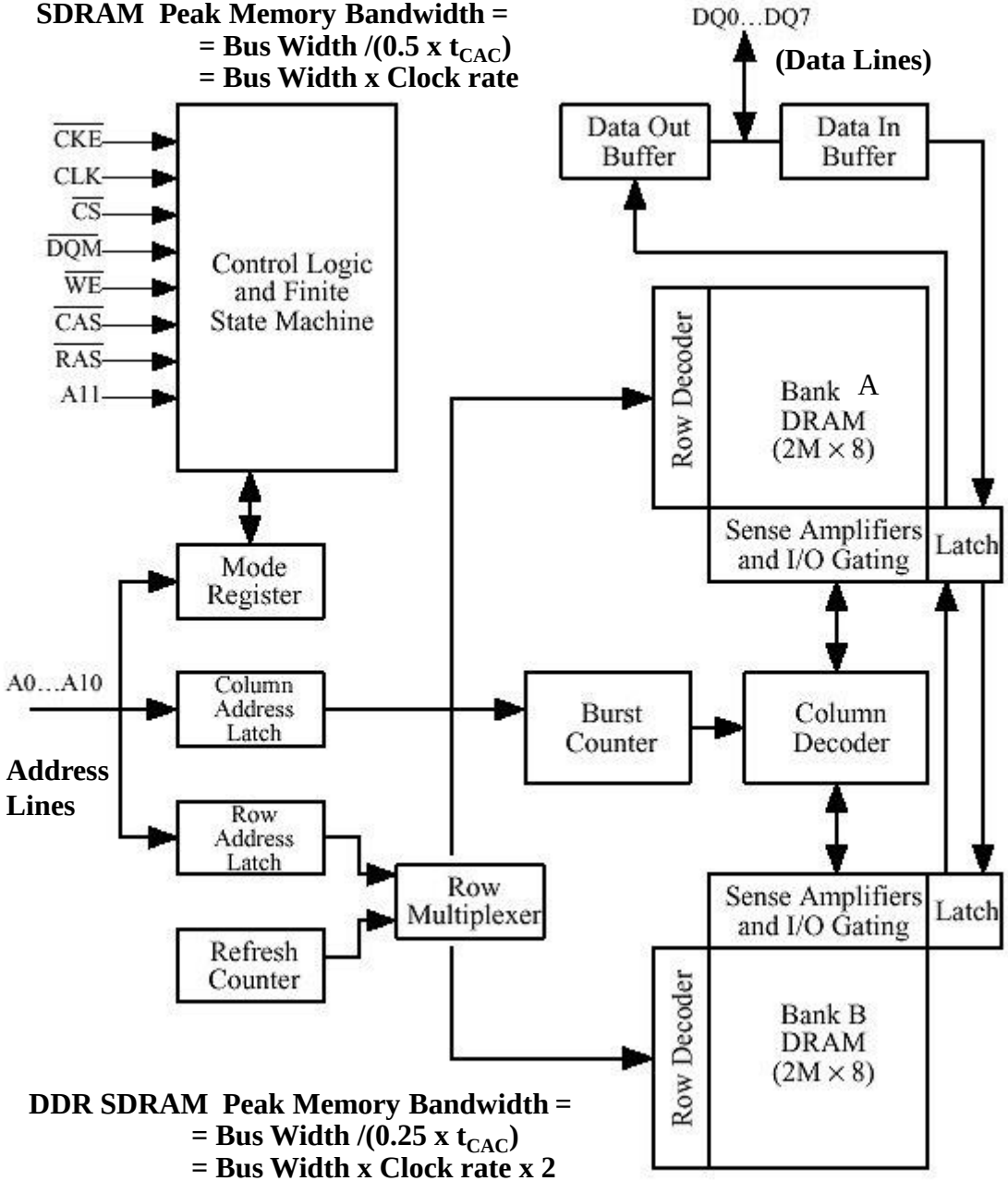
Bank Width = One Word
Bank Number = (Word Address) Mod (4)

Memory Bank Interleaving



Bank interleaving ***does not reduce latency of accesses to the same bank***

SDRAM Peak Memory Bandwidth =
= Bus Width /(0.5 x t_{CAC})
= Bus Width x Clock rate



DDR SDRAM Peak Memory Bandwidth =
= Bus Width /(0.25 x t_{CAC})
= Bus Width x Clock rate x 2

Synchronous Dynamic RAM, (SDRAM) (mid 90s) Organization

SDRAM speed is rated at max. clock speed supported:
100MHZ = PC100
133MHZ = PC133

DDR SDRAM

(late 90s - current)

organization is similar but **<u>four banks</u>** are used in each DDR SDRAM chip instead of two.

Data transfer on both **<u>rising and falling edges of the clock</u>**

DDR SDRAM rated by maximum memory bandwidth
PC3200 = 8 bytes x 200 MHz x 2
= 3200 Mbytes/sec

DDR SDRAM

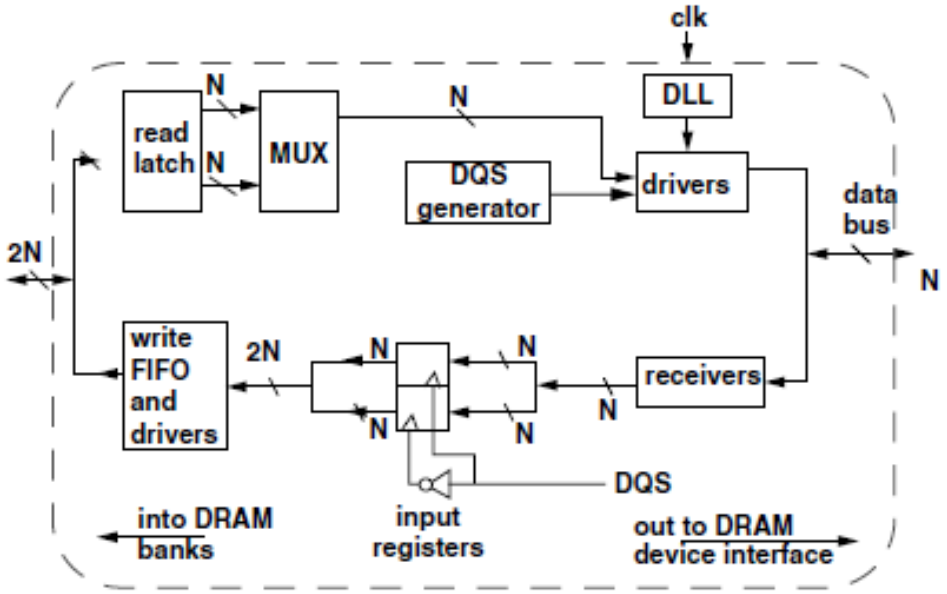
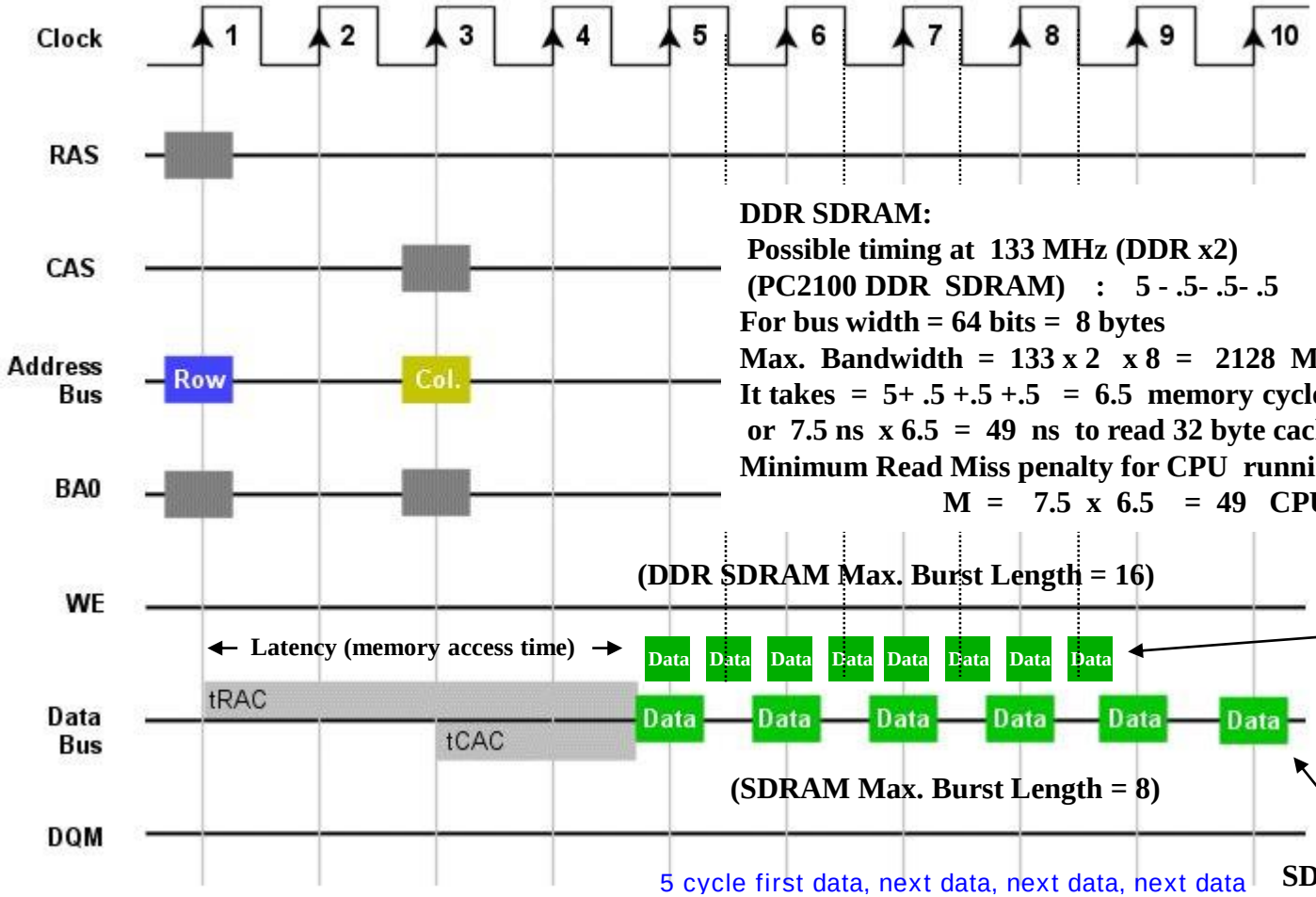


FIGURE 12.17: DDR SDRAM device I/O.

- **M-bit prefetch architecture**
  - Data rate multiplication architecture
  - M represents multiplication factor b/t DRAM device's internal width of data movement and width of the data bus on the device interface
  - DDR (DDR II, DDR III) SDRAM supports 2-bit (4-bit, 8-bit) prefetch architecture

64byte into DRAM
8byte out to DRAM

Simplified SDRAM/DDR SDRAM Read Timing



Twice as fast as SDRAM?

DDR SDRAM:
Possible timing at 133 MHz (DDR x2)
(PC2100 DDR SDRAM) : 5 - .5- .5- .5
For bus width = 64 bits = 8 bytes
Max. Bandwidth = 133 x 2 x 8 = 2128 Mbytes/sec
It takes = 5+ .5 +.5 +.5 = 6.5 memory cycles
or 7.5 ns x 6.5 = 49 ns to read 32 byte cache block
Minimum Read Miss penalty for CPU running at 1 GHz =
M = 7.5 x 6.5 = 49 CPU cycles

(DDR SDRAM Max. Burst Length = 16)

DDR SDRAM (Late 90s-Now)

(SDRAM Max. Burst Length = 8)

5 cycle first data, next data, next data, next data

SDRAM (mid 90s)

SDRAM Typical timing at 133 MHz (PC133 SDRAM) : 5-1-1-1
For bus width = 64 bits = 8 bytes Max. Bandwidth = 133 x 8 = 1064 Mbytes/sec
It takes = 5+1+1+1 = 8 memory cycles or 7.5 ns x 8 = 60 ns to read 32 byte cache block
Minimum Read Miss penalty for CPU running at 1 GHz = M = 7.5 x 8 = 60 CPU cycles

In this example for SDRAM: M = 60 cycles for DDR SDRAM: M = 49 cycles
Thus accounting for access latency DDR is 60/49 = 1.22 times faster
Not twice as fast (2128/1064 = 2) as indicated by peak bandwidth!

Memory Width, Interleaving: Performance Example

Given the following system parameters with single unified cache level L_1 (ignoring write policy):

Block size= 1 word Memory bus width= 1 word Miss rate =3% M = Miss penalty = 32 cycles

(4 cycles to send address 24 cycles access time, 4 cycles to send a word to CPU)

4 cycles	24 cycles	4 cycles

Miss Penalty = M= 4 + 24 + 4 = 32 cycles

(Base system)

Memory access/instruction = 1.2 $CPI_{execution}$ (ignoring cache misses) = 2

Miss rate (block size = 2 word = 8 bytes) = 2% Miss rate (block size = 4 words = 16 bytes) = 1%

- **The CPI of the base machine with 1-word blocks = 2 + (1.2 x 0.03 x 32) = 3.15**

Increasing the block size to two words (64 bits) gives the following CPI: (miss rate = 2%)

- **32-bit bus and memory, no interleaving, M = 2 x 32 = 64 cycles CPI = 2 + (1.2 x .02 x 64) = 3.54**
- **32-bit bus and memory, interleaved, M = 4 + 24 + 8 = 36 cycles CPI = 2 + (1.2 x .02 x 36) = 2.86**
- **64-bit bus and memory, no interleaving, M = 32 cycles CPI = 2 + (1.2 x 0.02 x 32) = 2.77**

Increasing the block size to four words (128 bits); resulting CPI: (miss rate = 1%)

- **32-bit bus and memory, no interleaving , M = 4 x 32 = 128 cycles CPI = 2 + (1.2 x 0.01 x 128) = 3.54**
- **32-bit bus and memory, interleaved , M = 4 + 24 + 16 = 44 cycles CPI = 2 + (1.2 x 0.01 x 44) = 2.53**
- **64-bit bus and memory, no interleaving, M = 2 x 32 = 64 cycles CPI = 2 + (1.2 x 0.01 x 64) = 2.77**
- **64-bit bus and memory, interleaved, M = 4 + 24 + 8 = 36 cycles CPI = 2 + (1.2 x 0.01 x 36) = 2.43**
- **128-bit bus and memory, no interleaving, M = 32 cycles CPI = 2 + (1.2 x 0.01 x 32) = 2.38**

Miss Penalty = M = Number of CPU stall cycles for an access missed in cache and satisfied by main memory