N26F300
VLSI SYSTEM DESIGN
(GRADUATE LEVEL)

Memory Subsystems: Dynamic Random Memories (DRAMs)

Outline

- DRAM Overview
- DRAM Device and Structure
- DRAM Subsystem Organization

Reference: “Memory Systems Cache, DRAM, DISK” by Bruce Jacob, Spencer W. Ng and David T. Wang

[part of slides are adopted from MOE materials “DRAM Sub-system,” by Prof. CT Huang, NTHU, 2013]

DRAM Overview

Why DRAM matters?

- Patented in 1968 by Dennard
- Significantly cheaper than SRAM
  - Higher density than SRAMs
    - 1T vs. 6T
  - A bit is represented by a high or low charge on the capacitor
- Significantly slower than SRAM
  - Off-chip (external) vs. on-chip (embedded)
- Disadvantages
  - Longer access times
  - Leaky, needs to be refreshed
  - Cannot be easily integrated with CMOS

Basics of DRAM

DRAM

- DRAM (Dynamic RAM)
- Used mostly in main mem.
- Capacitor + 1 transistor/bit
- Need refresh every 4-8 ms
  - 5% of total time
- Read is destructive (need for write-back)
- Access time < cycle time (because of writing back)
- Density (25-50):1 to SRAM
- Address lines multiplexed
  - pins are scarce!

SRAM

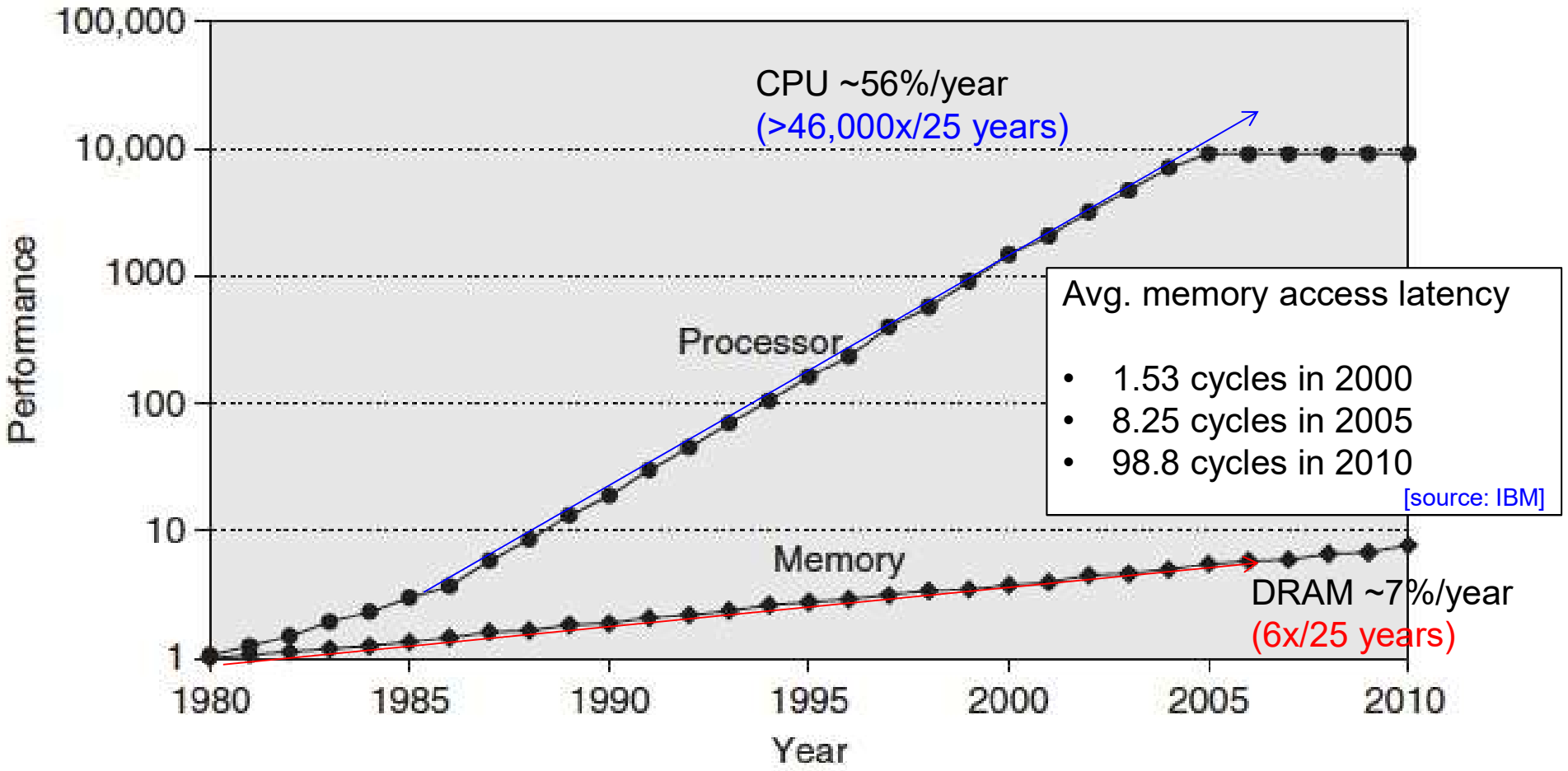
- SRAM (Static RAM)
- Used mostly in caches (I, D, TLB, BTB)
- 1 flip-flop (4-6 transistors) per bit
- Read is not destructive
- Access time = cycle time
- Speed (8-16):1 to DRAM
- Address lines not multiplexed
  - high speed of decoding imp.

Cost of Different Kinds of Storage Devices

Technology	Typical Access	$ per GB in 2008
SRAM	0.5 – 2.5 ns	$2000 - $5000
DRAM	50 – 70 ns	$20 - $75
Magnetic Disk	5 – 20 ms	$0.2 - $2

Processor-DRAM Memory Gap

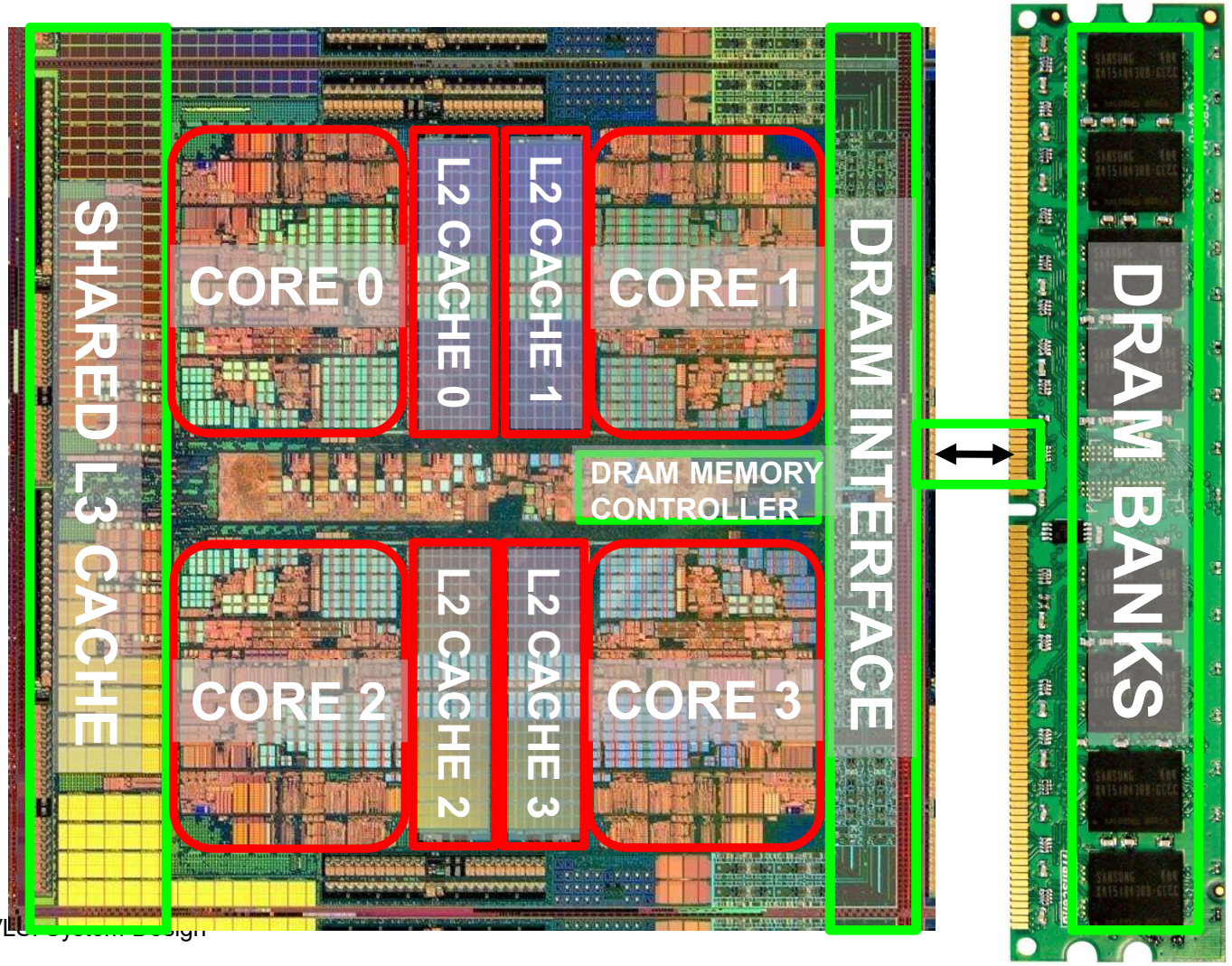
Processor-DRAM Performance Gap grows ~50% / year



Memory Issues

- Latency
  - Time to move through the longest circuit path (from the start of request to the response)
- Bandwidth
  - Number of bits transported at one time
- Capacity
  - Size of memory
- Energy
  - Cost of accessing memory (to read and write)

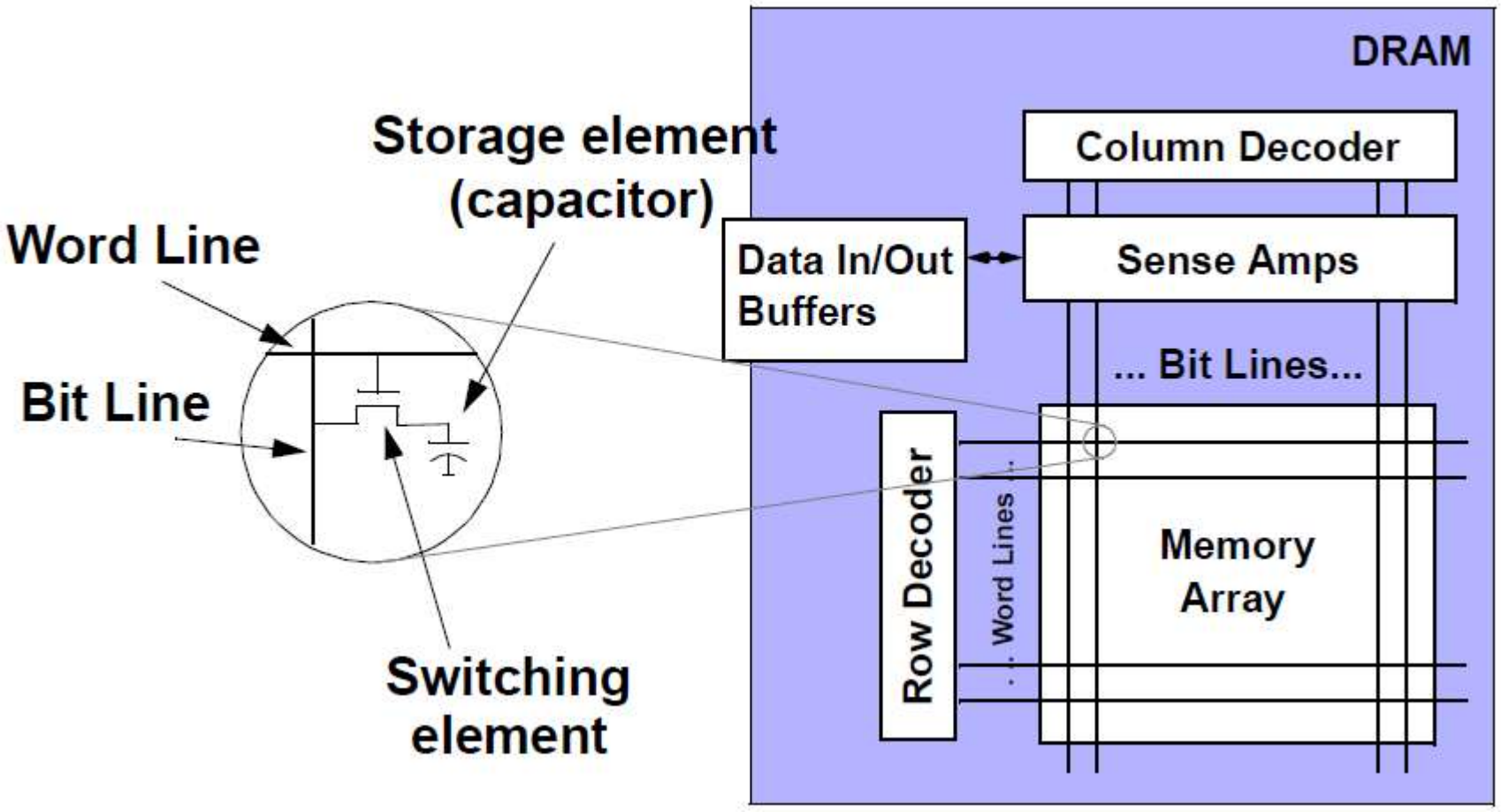
Main Memory in The System



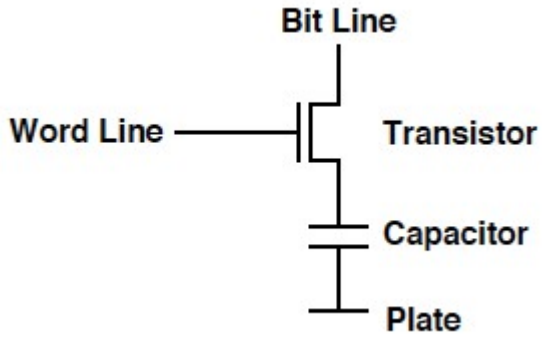
10

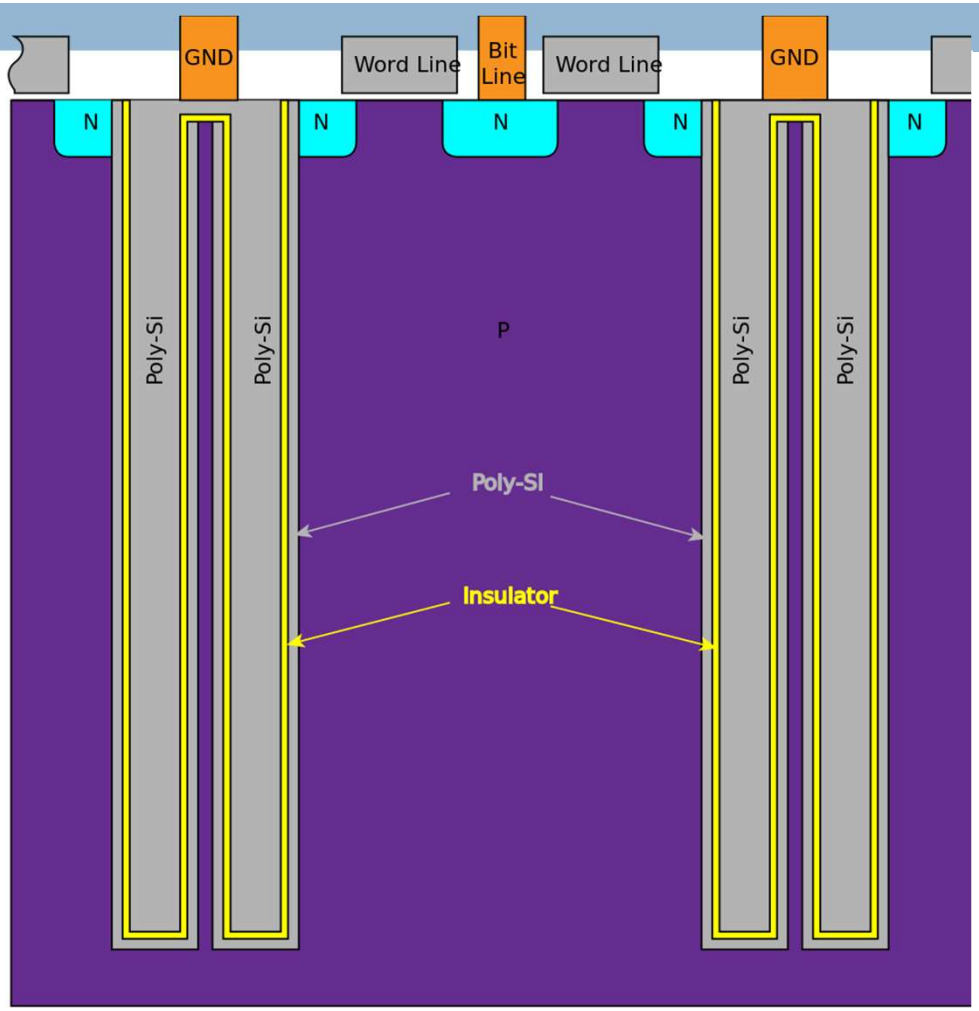
DRAM Device & Structure

DRAM Cell


DRAM
Storage element
(capacitor)
Word Line
Bit Line
Switching
element
Column Decoder
Data In/Out
Buffers
Sense Amps
... Bit Lines...
Row Decoder
. . . Word Lines . . .
Memory
Array

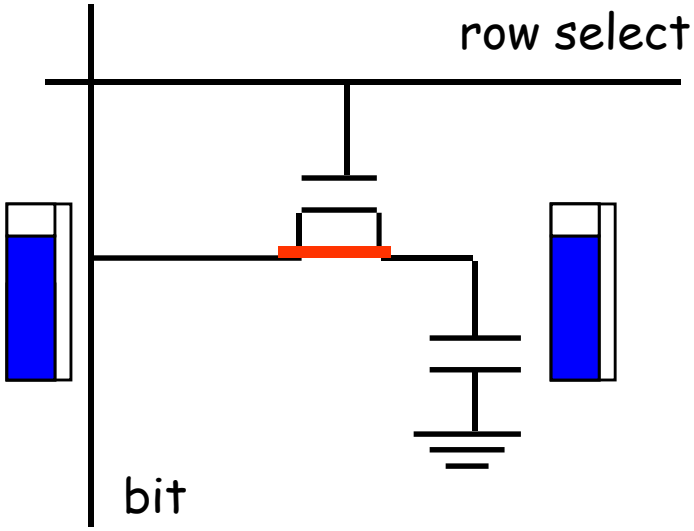
Trench DRAM Cell





1-Transistor Memory Cell (DRAM)

- Write:
  - 1. Drive bit line
  - 2. Select row
- Read:
  - 1. Precharge bit line to Vdd/2
  - 2. Select row
  - 3. Cell and bit line share charges
    - Minute voltage changes on the bit line
  - 4. Sense (fancy sense amp)
    - Can detect changes of ~1 million electrons
  - 5. Write: restore the value
- Refresh
  - 1. Just do a dummy read to every cell.

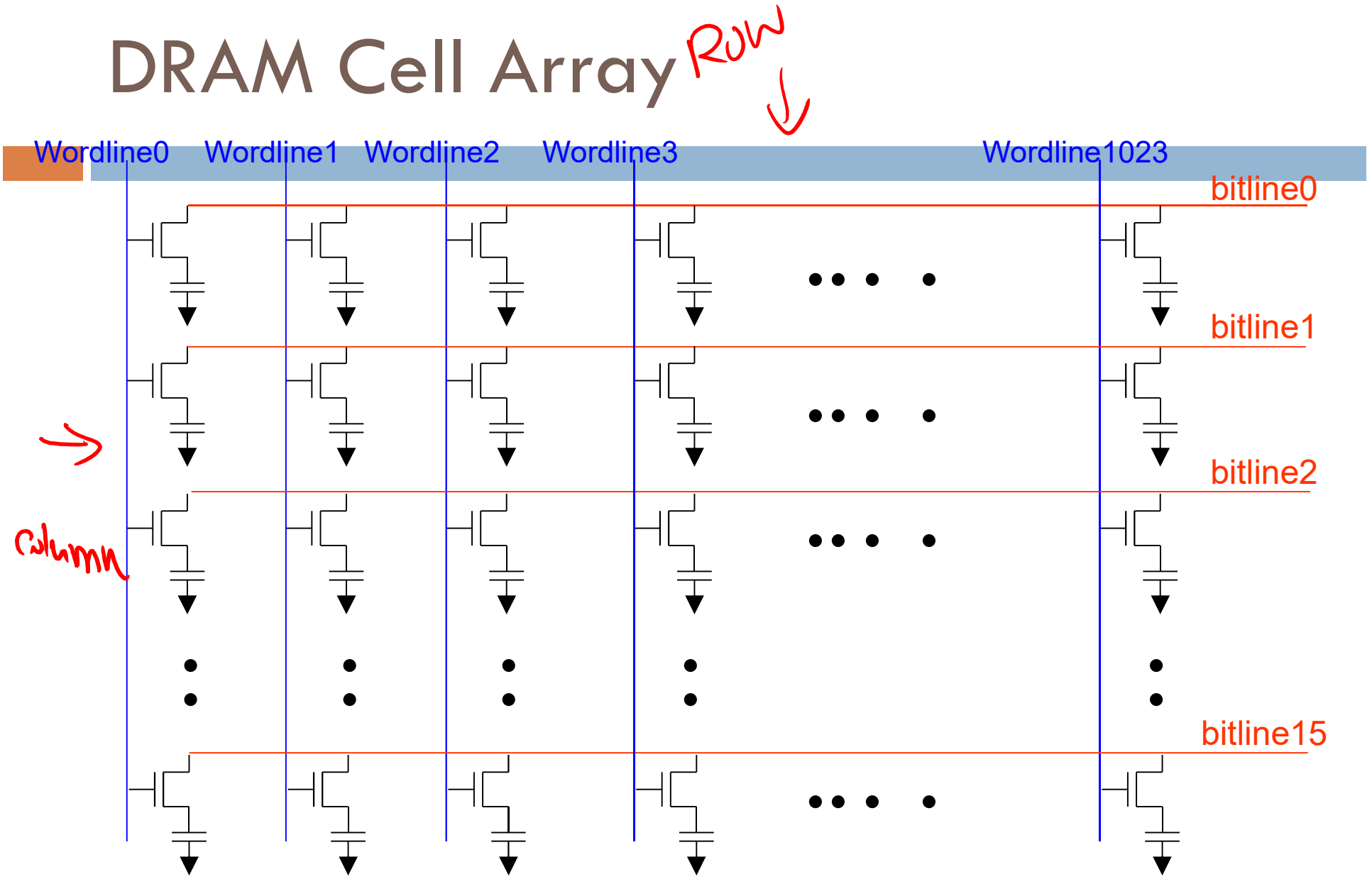


Read is really a read followed by a restoring write

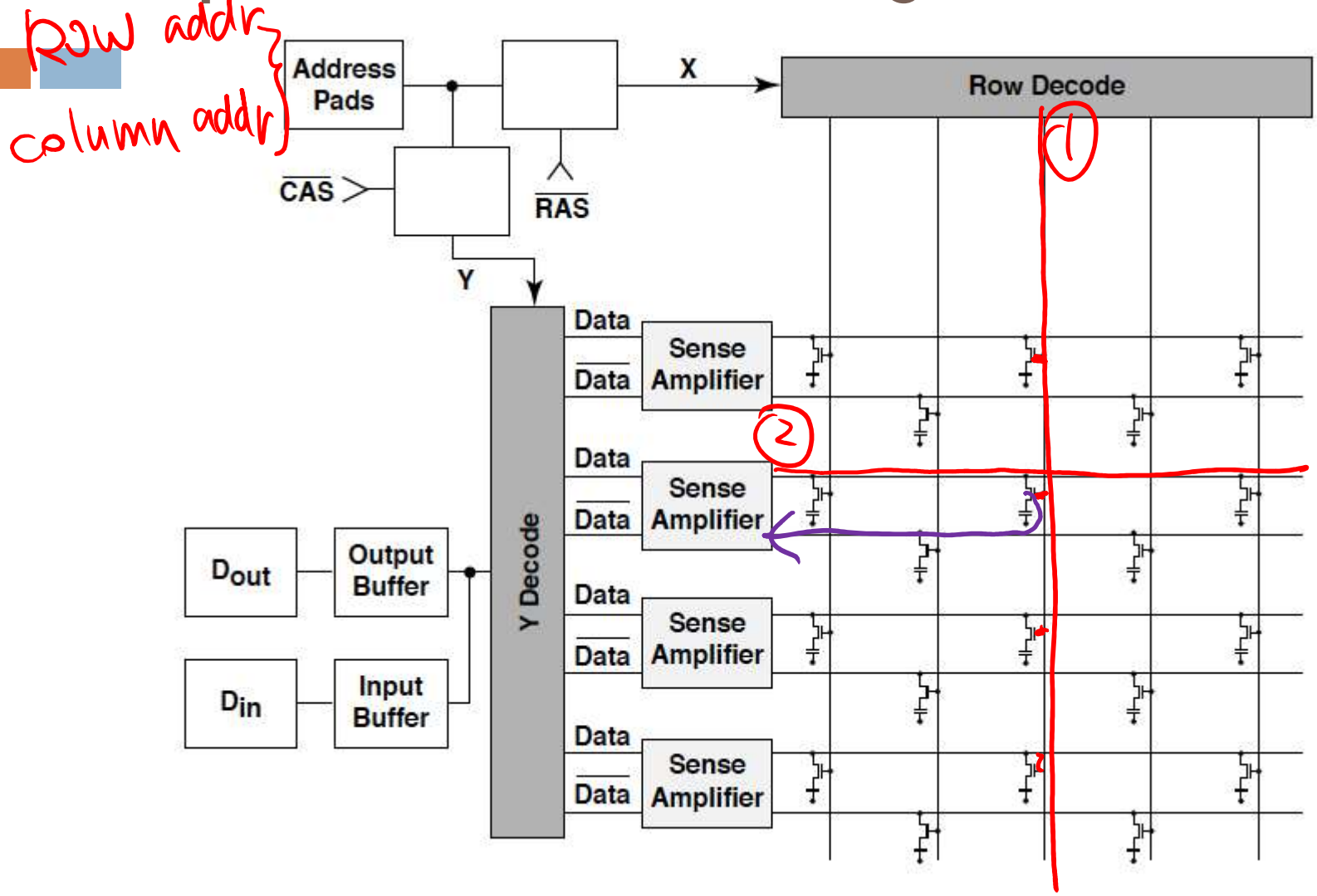
DRAM Cell Properties

- Voltage swing on bitline is small
  - Design target:
    - Bitline capacitance as small as possible; bit cell capacitance as large as possible to increase charge transfer
- Read is destructive
  - Part of read cycle is used to restore level inside of bit cell capacitor
- Capacitor leaks
  - Periodical refresh is mandatory
- Noise sources in DRAM are word line to bit line coupling, bit line to bit line coupling

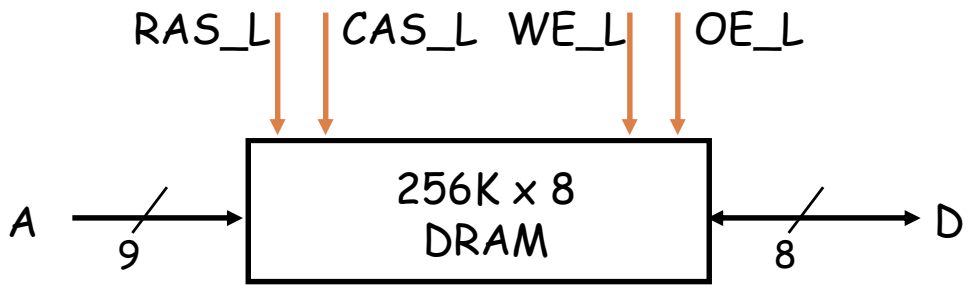
DRAM Cell Array



Simplified DRAM Diagram

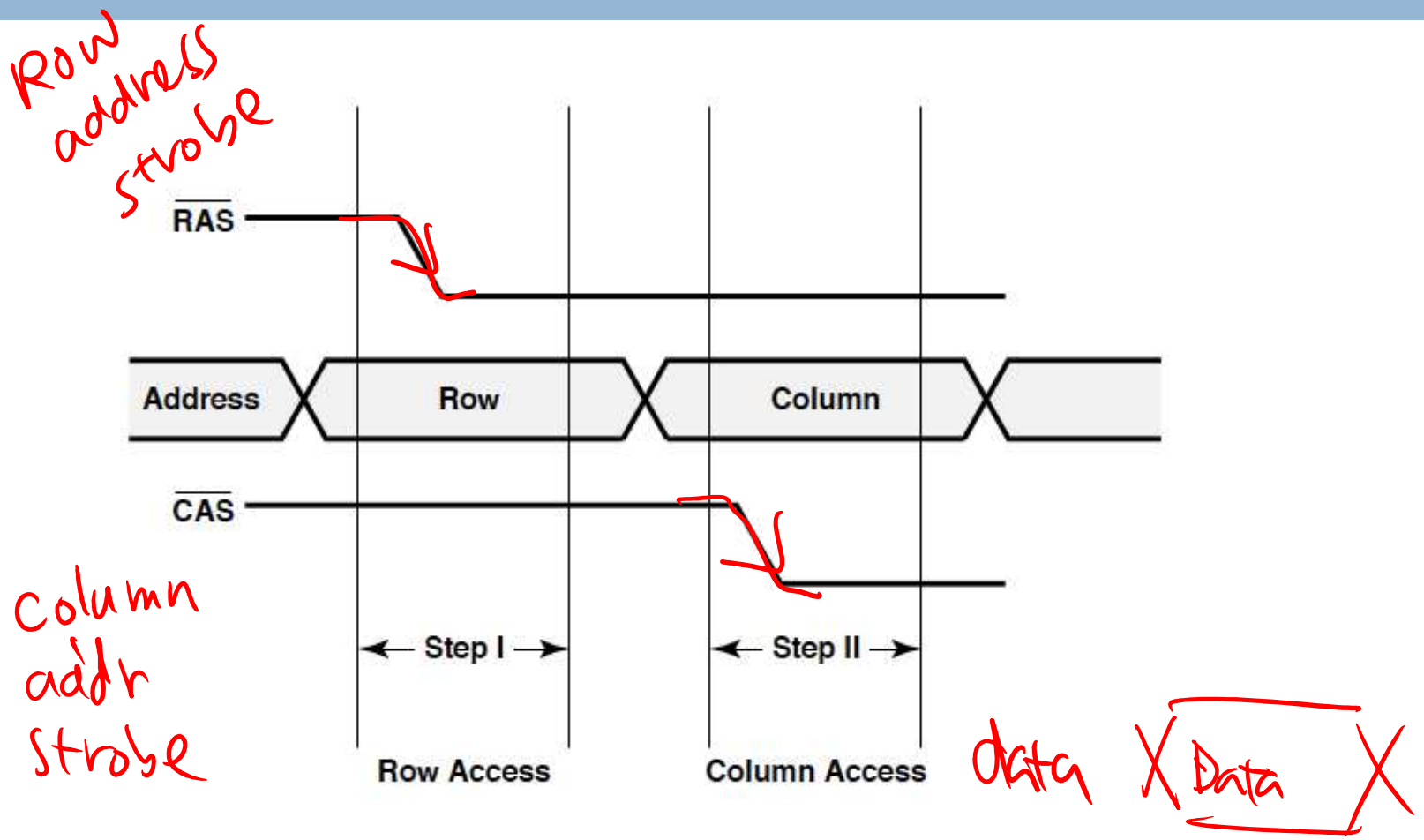


Logic Diagram of a Typical DRAM



- Control Signals (RAS_L, CAS_L, WE_L, OE_L) are all active low
- Din and Dout are combined (D):
  - WE_L is asserted (Low), OE_L is disasserted (High)
    - D serves as the data input pin
  - WE_L is disasserted (High), OE_L is asserted (Low)
    - D is the data output pin
- Row and column addresses share the same pins (A)
  - RAS_L goes low: Pins A are latched in as row address
  - CAS_L goes low: Pins A are latched in as column address
  - RAS/CAS edge-sensitive

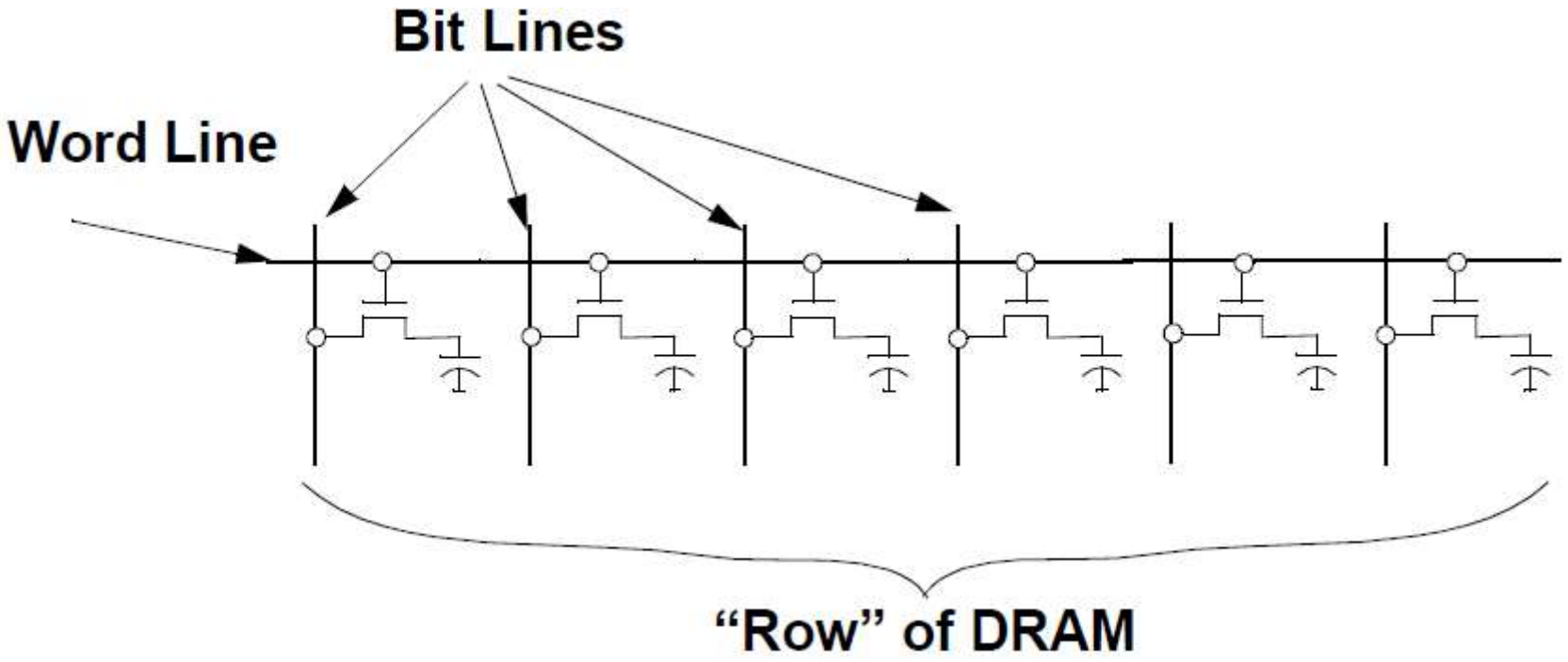
Simple DRAM Access Timing



Row

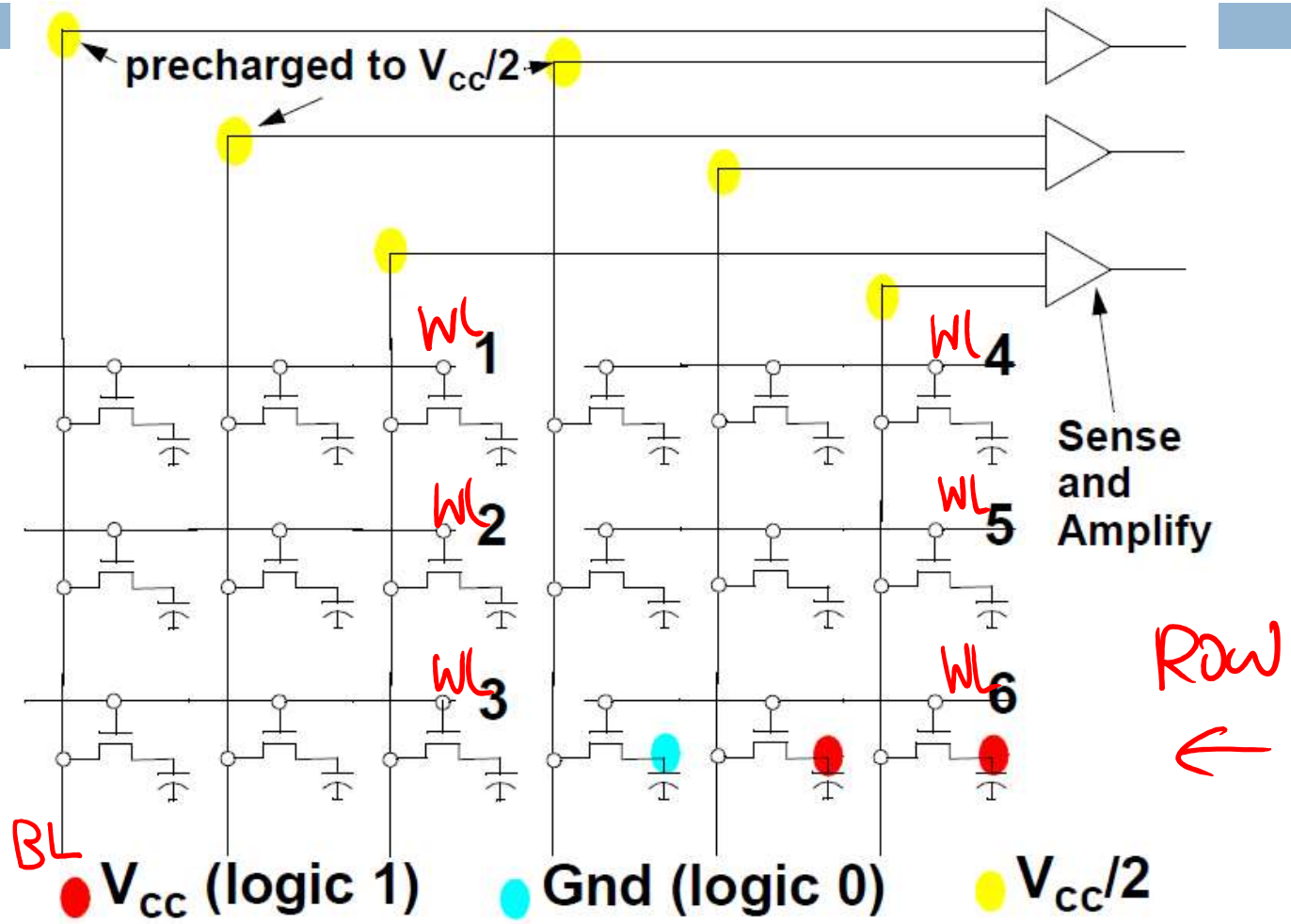
- Row Size: 8 Kb @ 256 Mb SDRAM(synchronous DRAM)

page

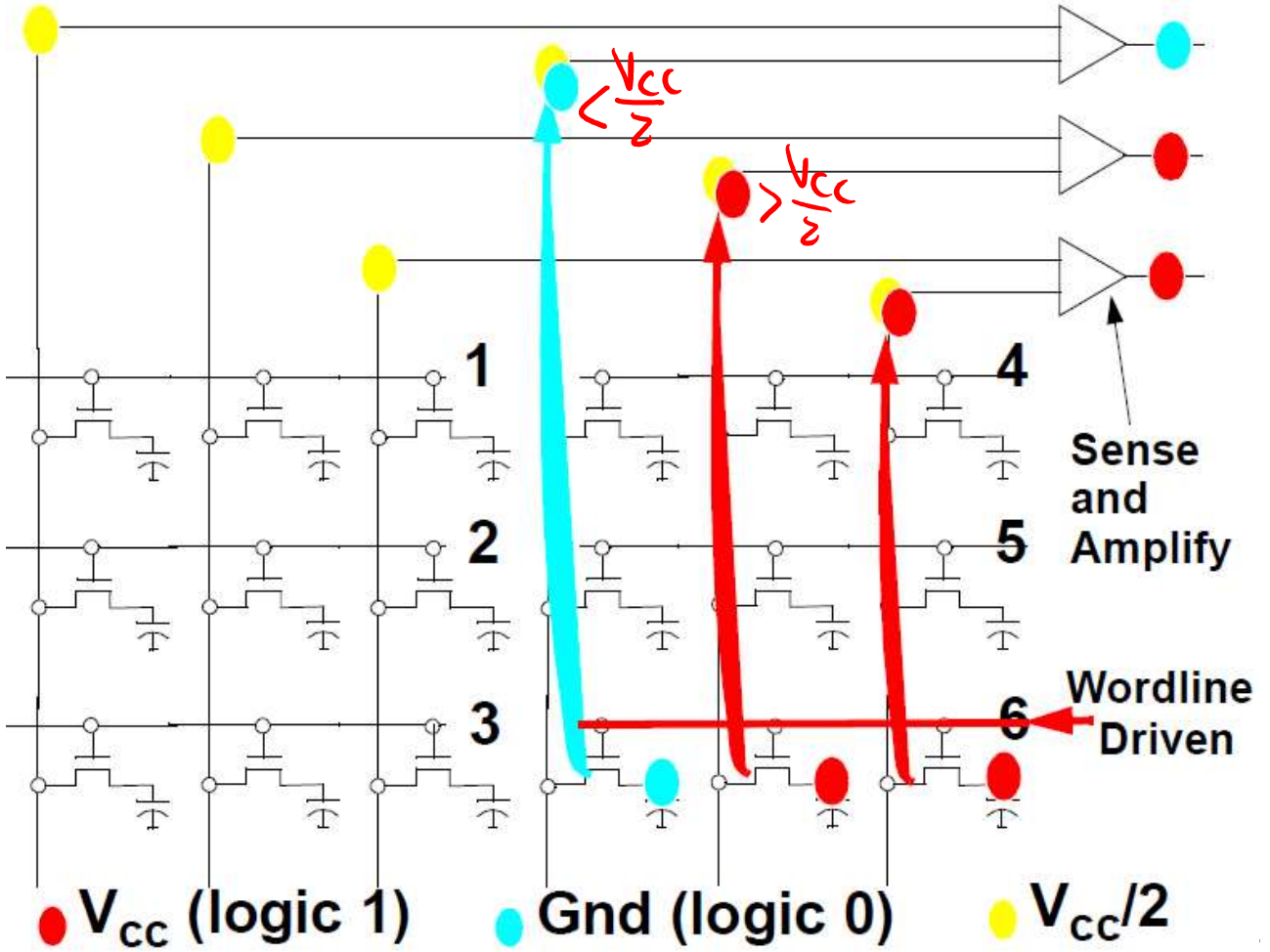


Page size: # of bits per row i.e., # of bits loaded into the sense Amps.

Precharge and Sense Amp

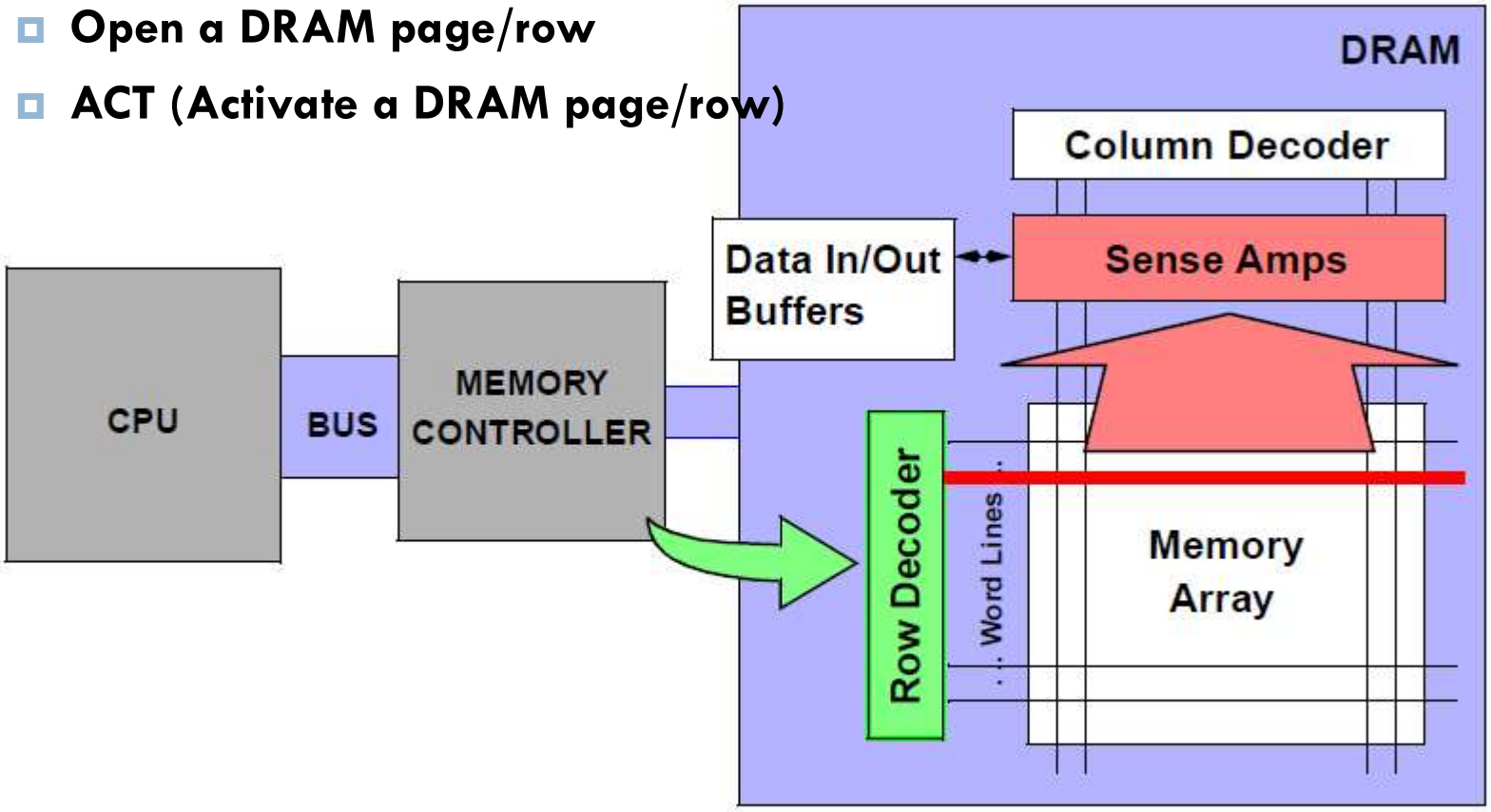


Destructive Read



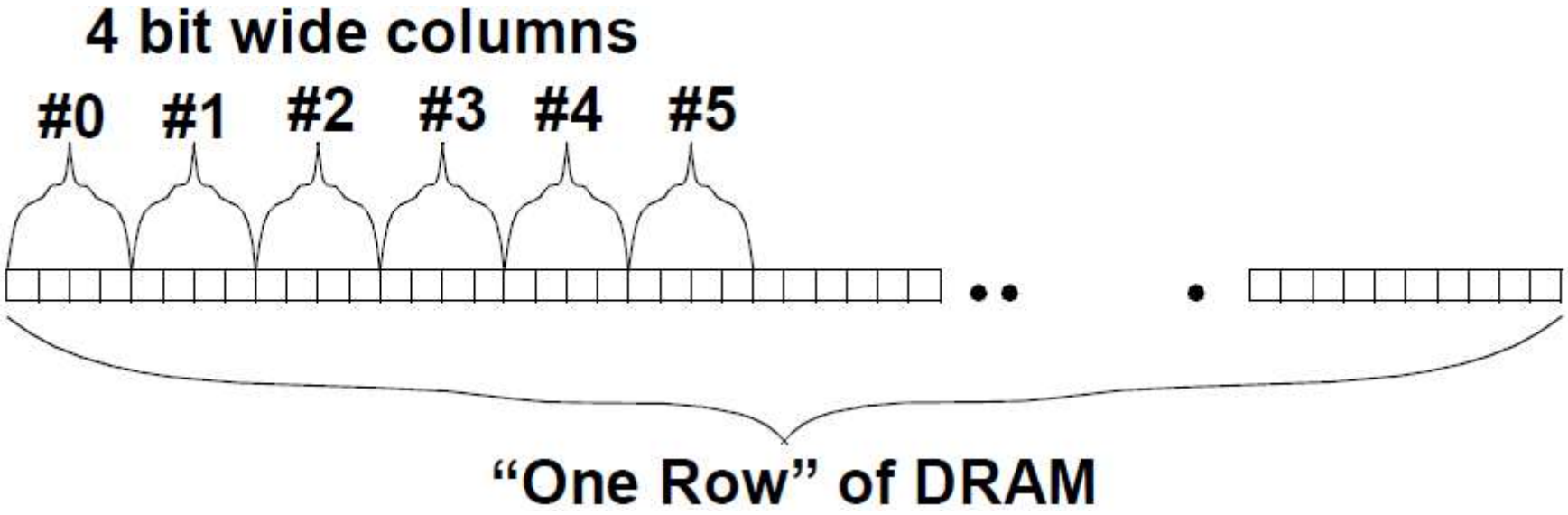
Row Access

- **RAS (Row Address Strobe)**
  - **Open a DRAM page/row**
  - **ACT (Activate a DRAM page/row)**

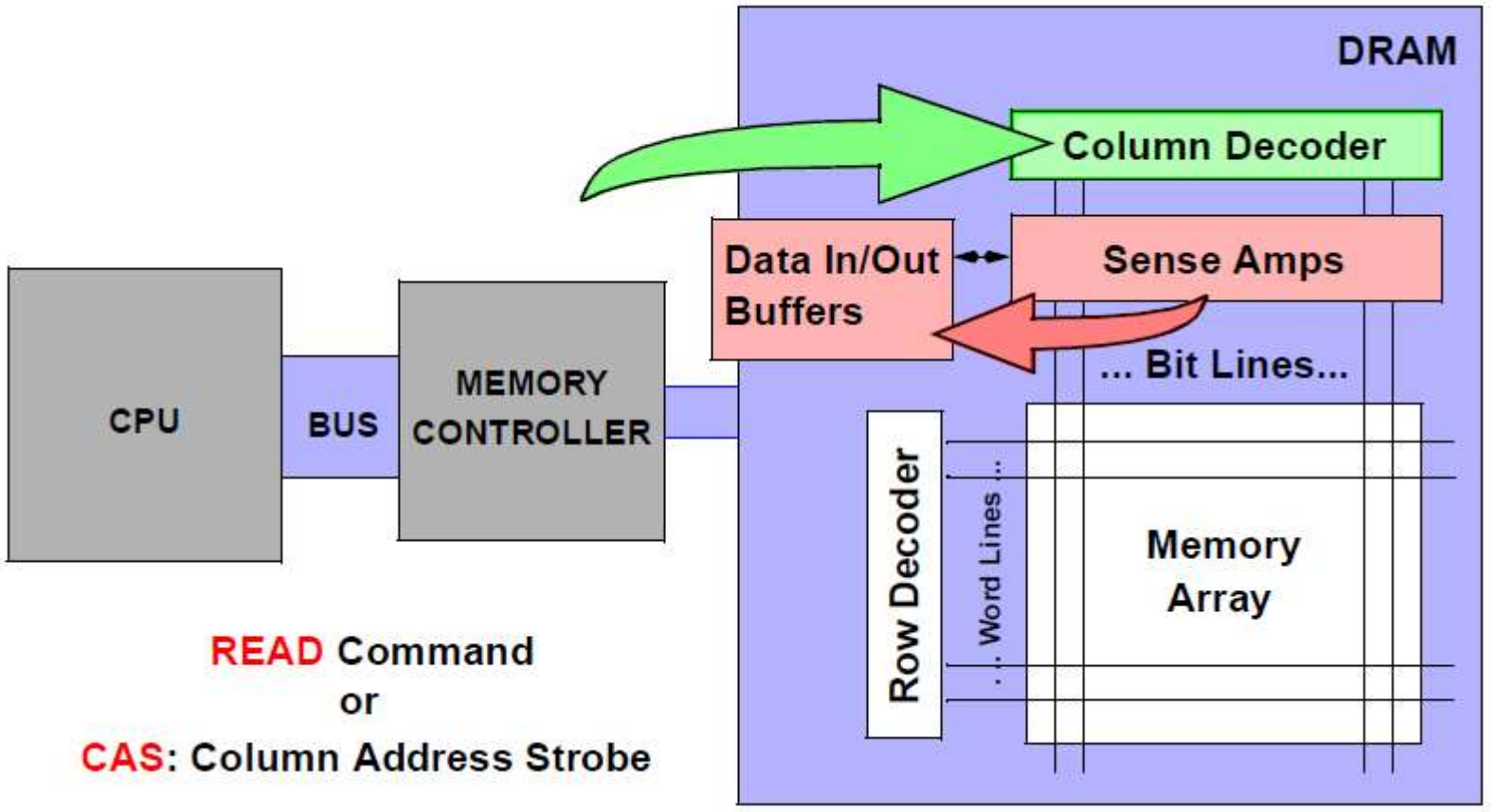


Column

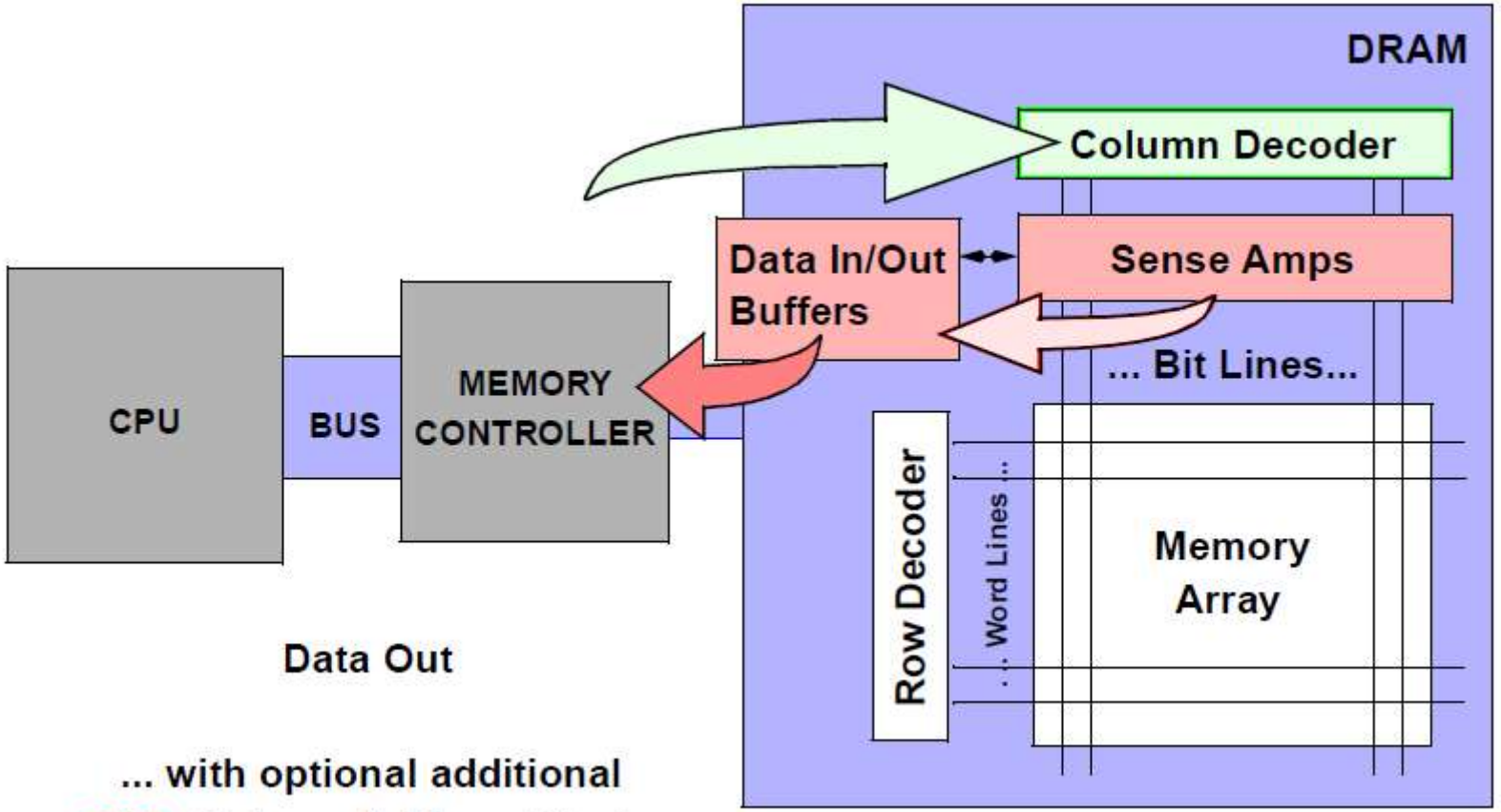
- Smallest addressable quantity of DRAM on chip
- For SDRAM (synchronous DRAM)
  - column size == chip data bus width (4, 8,16, 32)
  - get "n" columns per access. n = (1, 2, 4, 8)



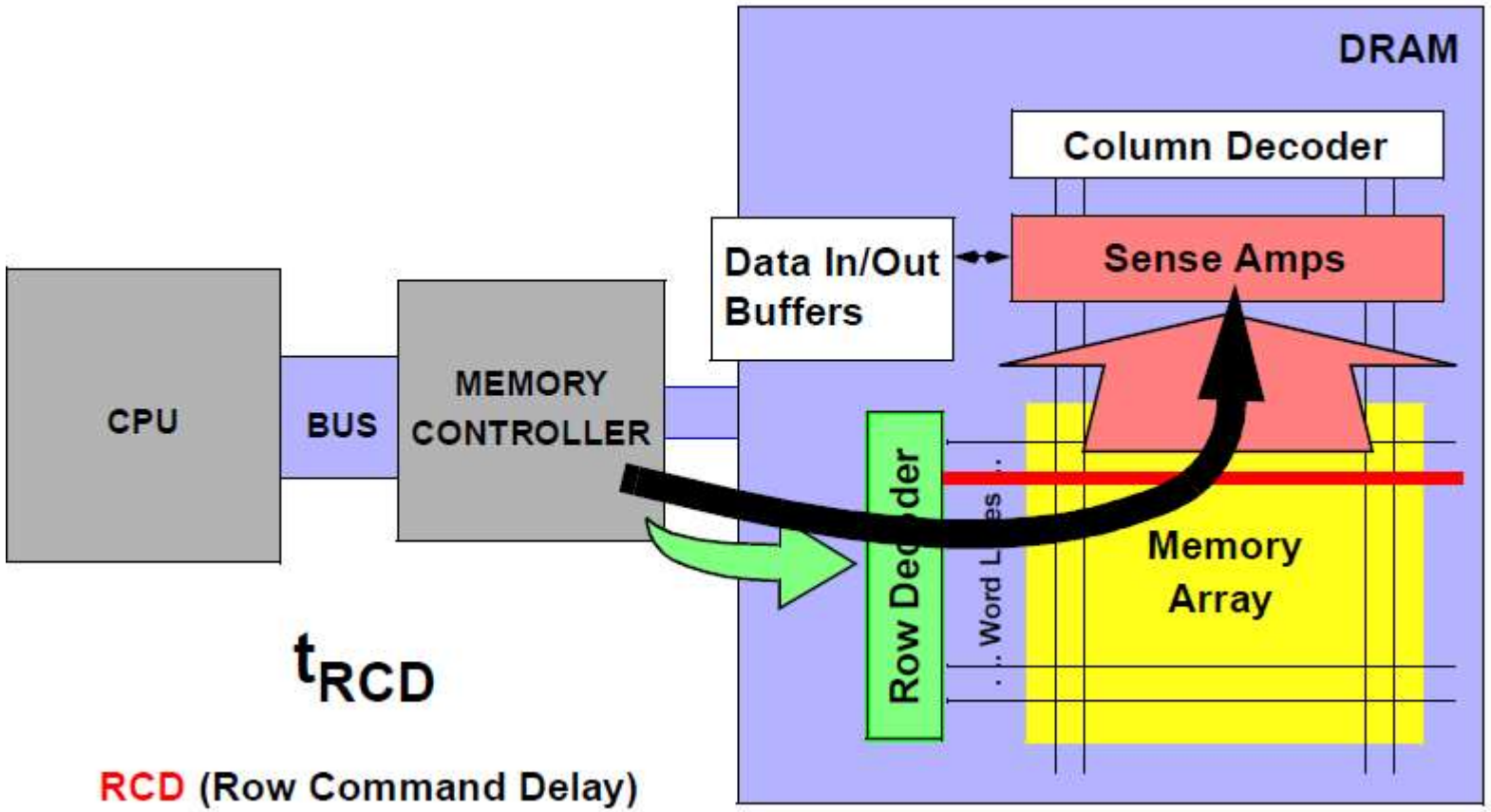
Column Access



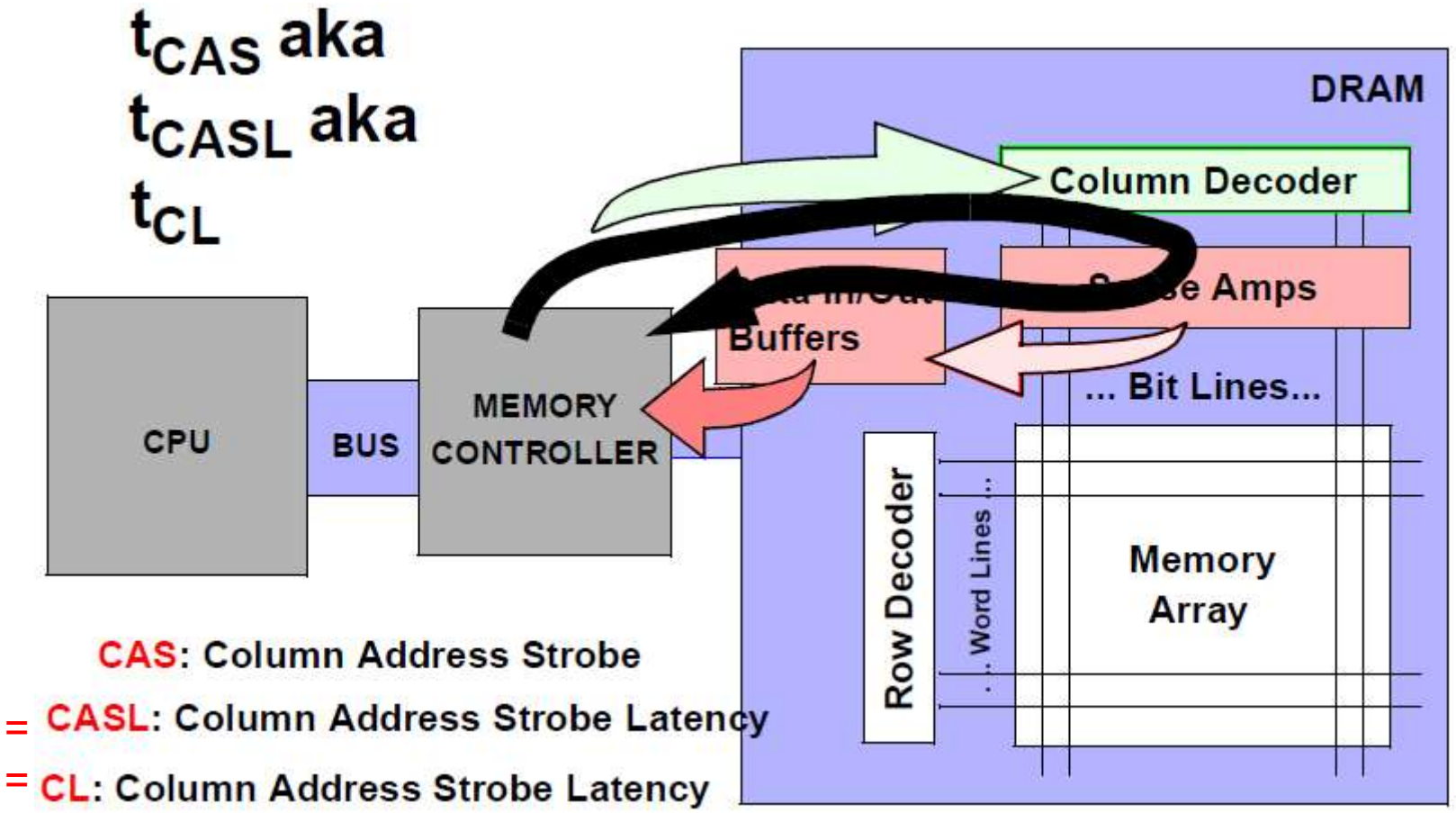
Data Out


DRAM
Column Decoder
Data In/Out Buffers
Sense Amps
... Bit Lines...
CPU
BUS
MEMORY CONTROLLER
Row Decoder
... Word Lines ...
Memory Array
Data Out
... with optional additional
CAS: Column Address Strobe

How fast can I move data from DRAM cell to sense amp?

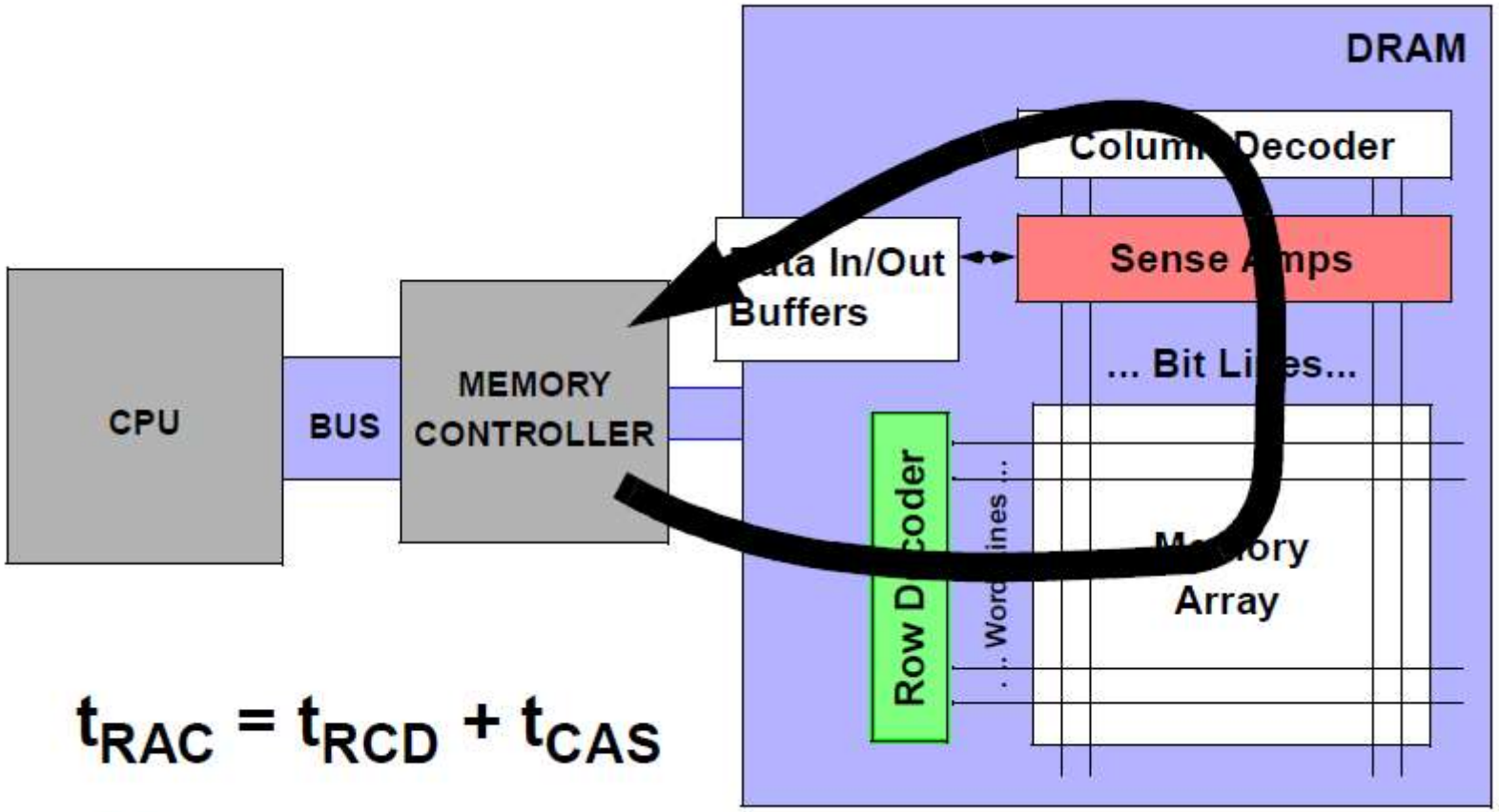


How fast can I get data out of sense amps into memory controller?



aka : as known as

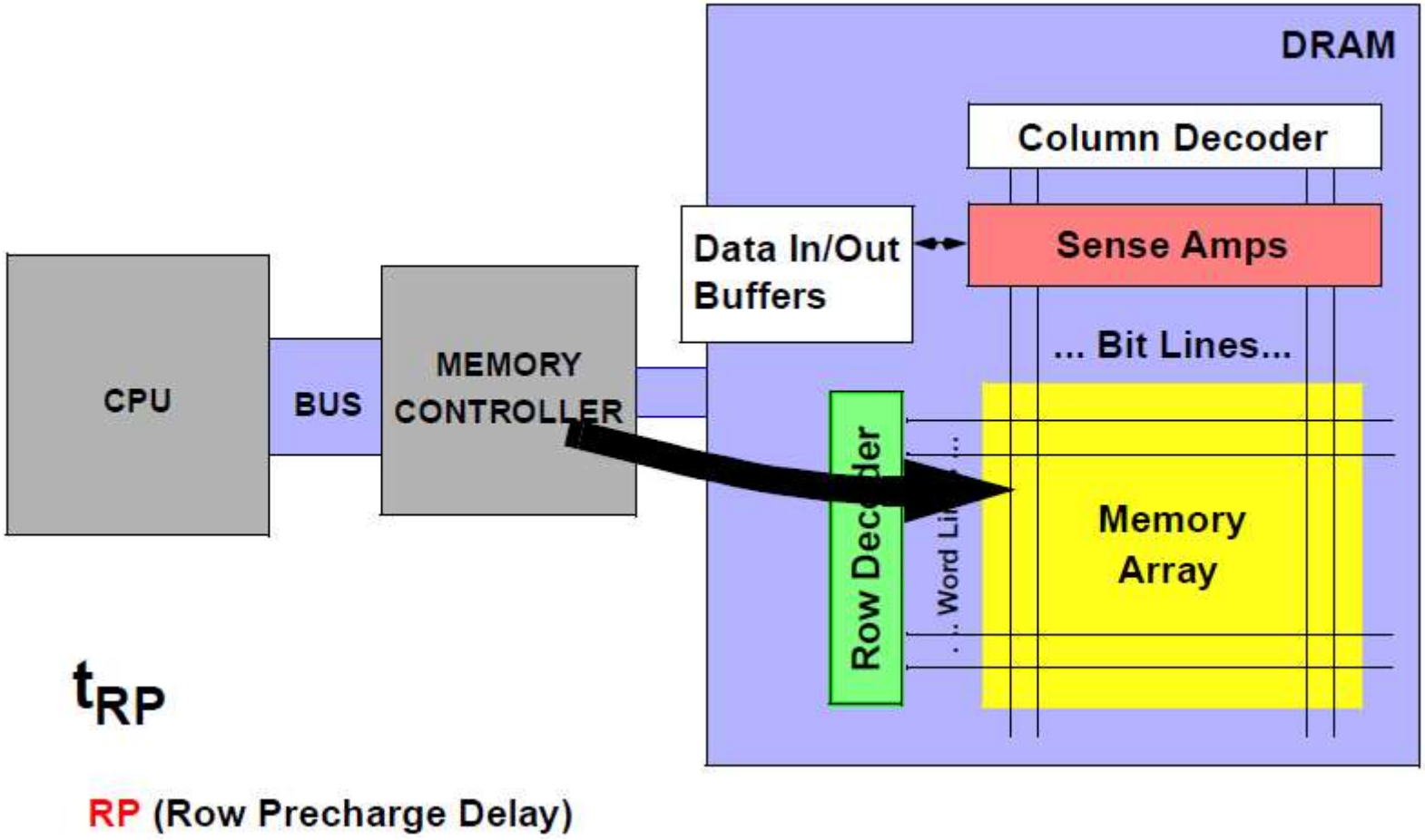
How fast can I move data from DRAM cell into memory controller?



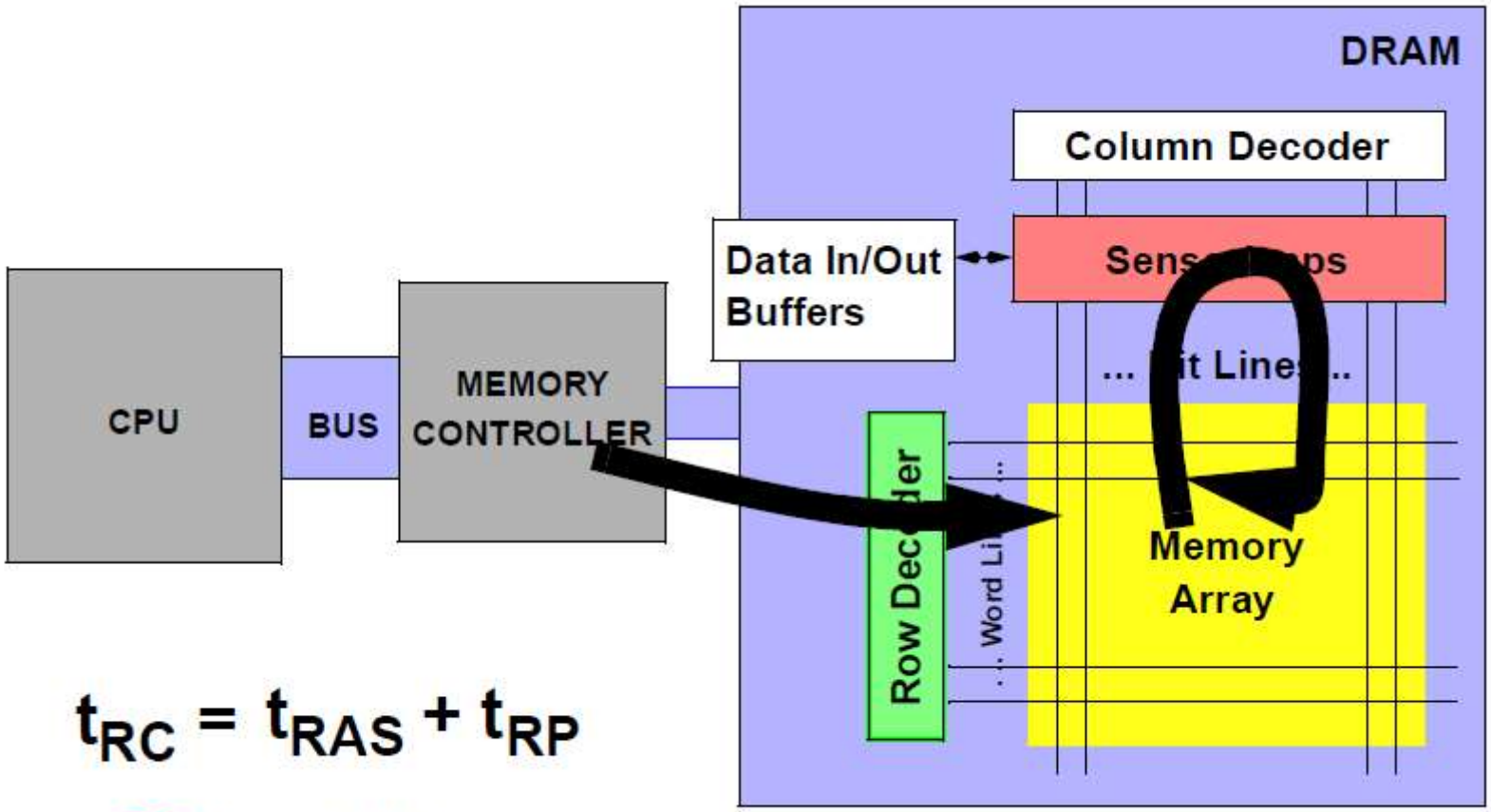
$$t_{RAC} = t_{RCD} + t_{CAS}$$

RAC (Random Access Delay)

How fast can I precharge DRAM array so I can engage another RAS?



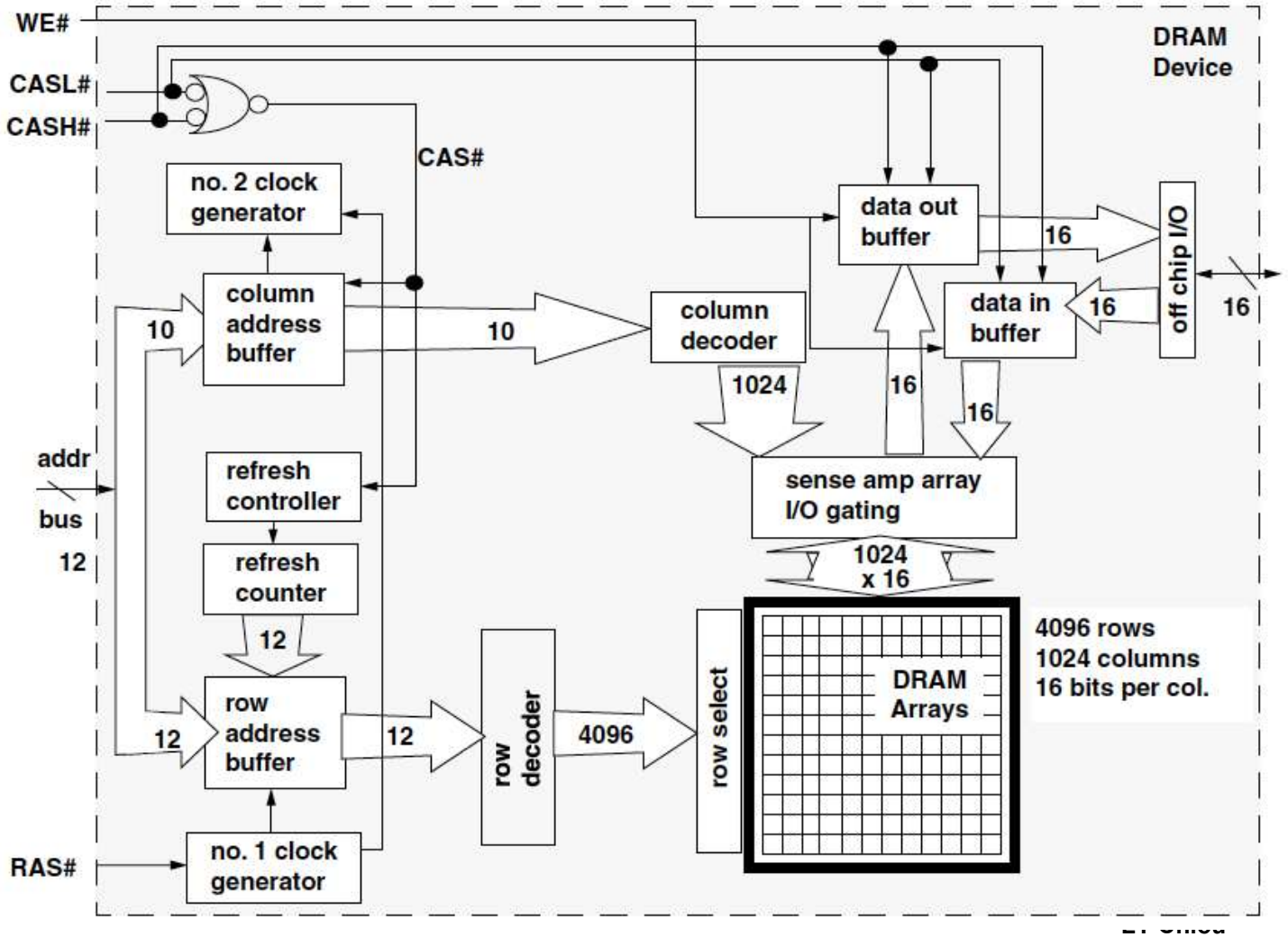
How fast can I read from different rows?



$$t_{RC} = t_{RAS} + t_{RP}$$

RC (Row Cycle Time)

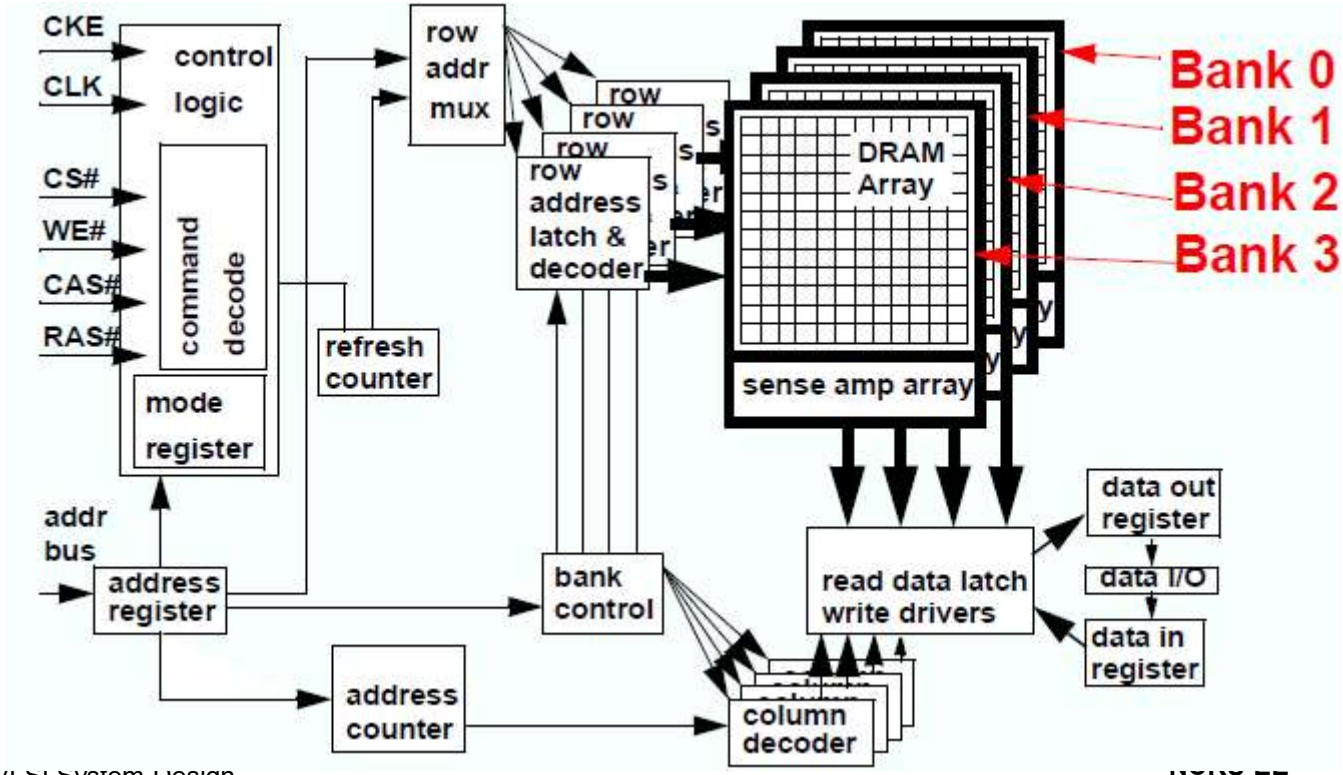
DRAM Organization



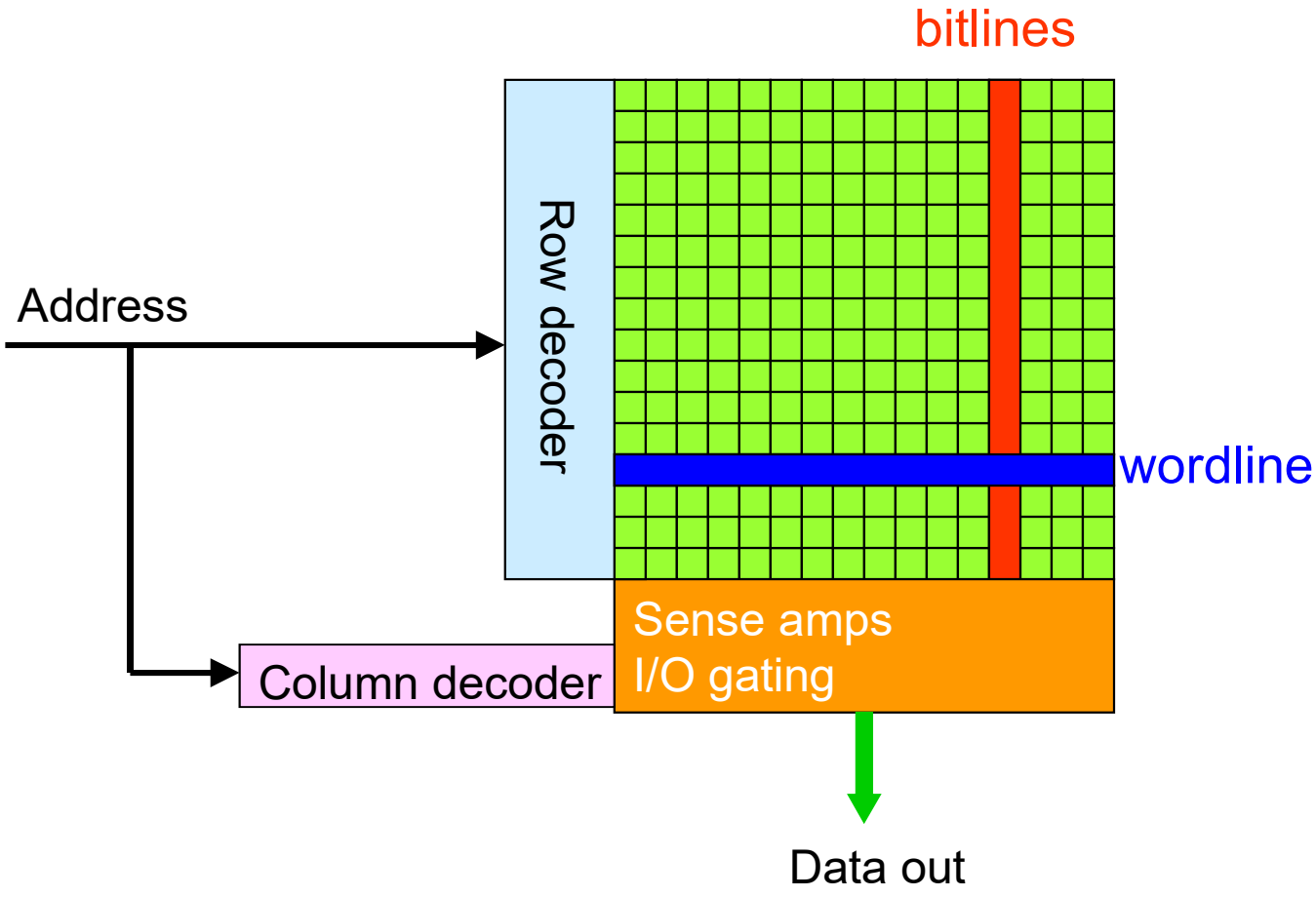
Source: Memory Systems Architecture Course, Bruce Jacob, Maryland

Bank

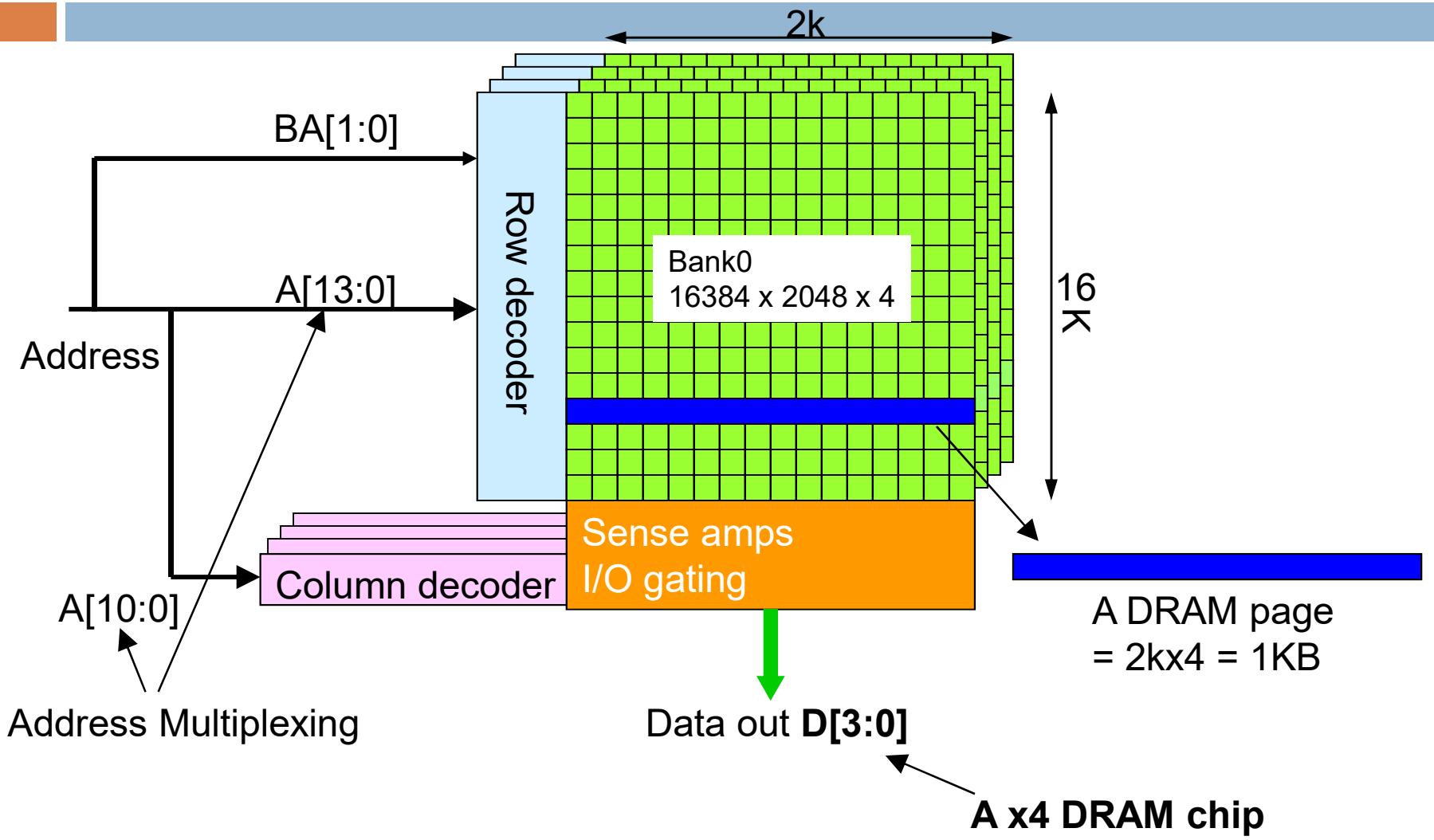
- "Banks" of independent memory arrays inside of a DRAM Chip



One DRAM Bank



Example: 512Mb 4-bank DRAM (x4)



DRAM Refresh

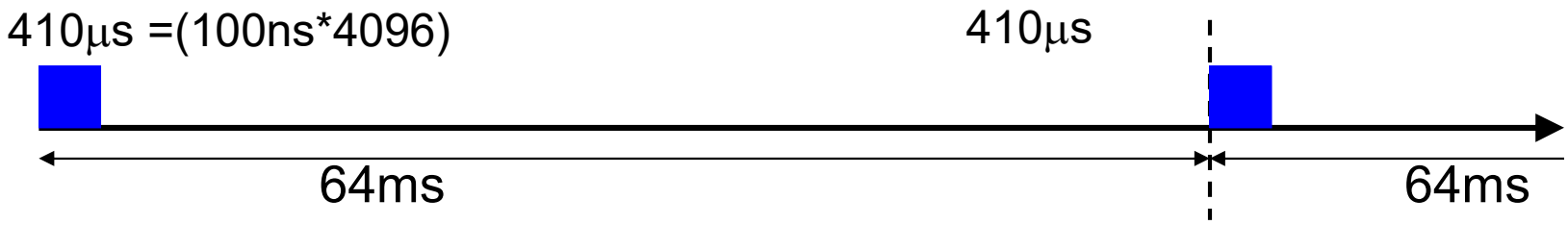
- Leaky storage
- Periodic Refresh across DRAM rows
- Un-accessible when refreshing
- Read, and write the same data back

- Example:
  - 4k rows in a DRAM
  - 100ns read cycle
  - Decay in 64ms

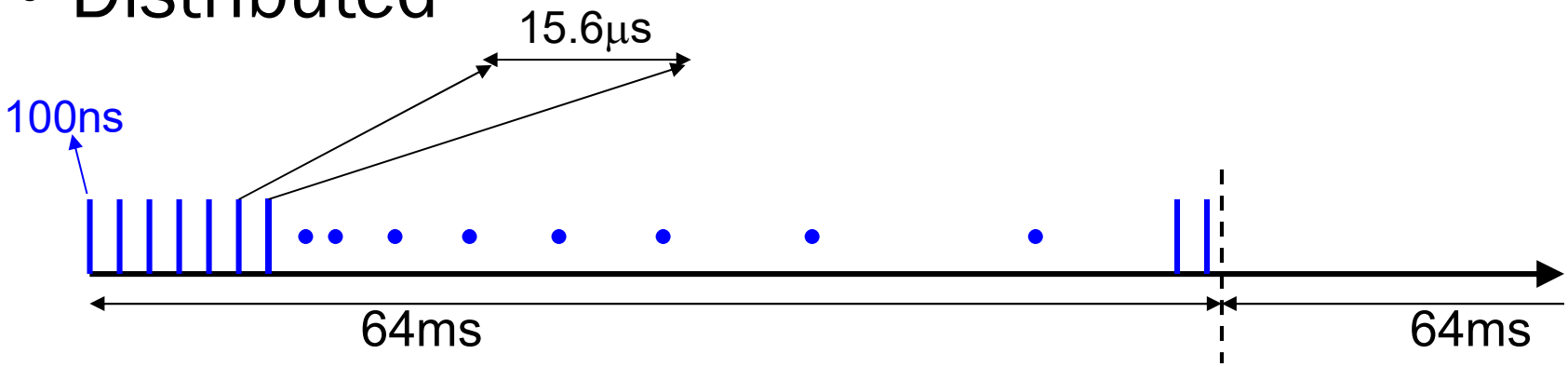
  - 4096*100ns ≅ 410μs to refresh once
  - 410μs / 64ms = 0.64% unavailability

DRAM Refresh Styles

- Bursty



- Distributed

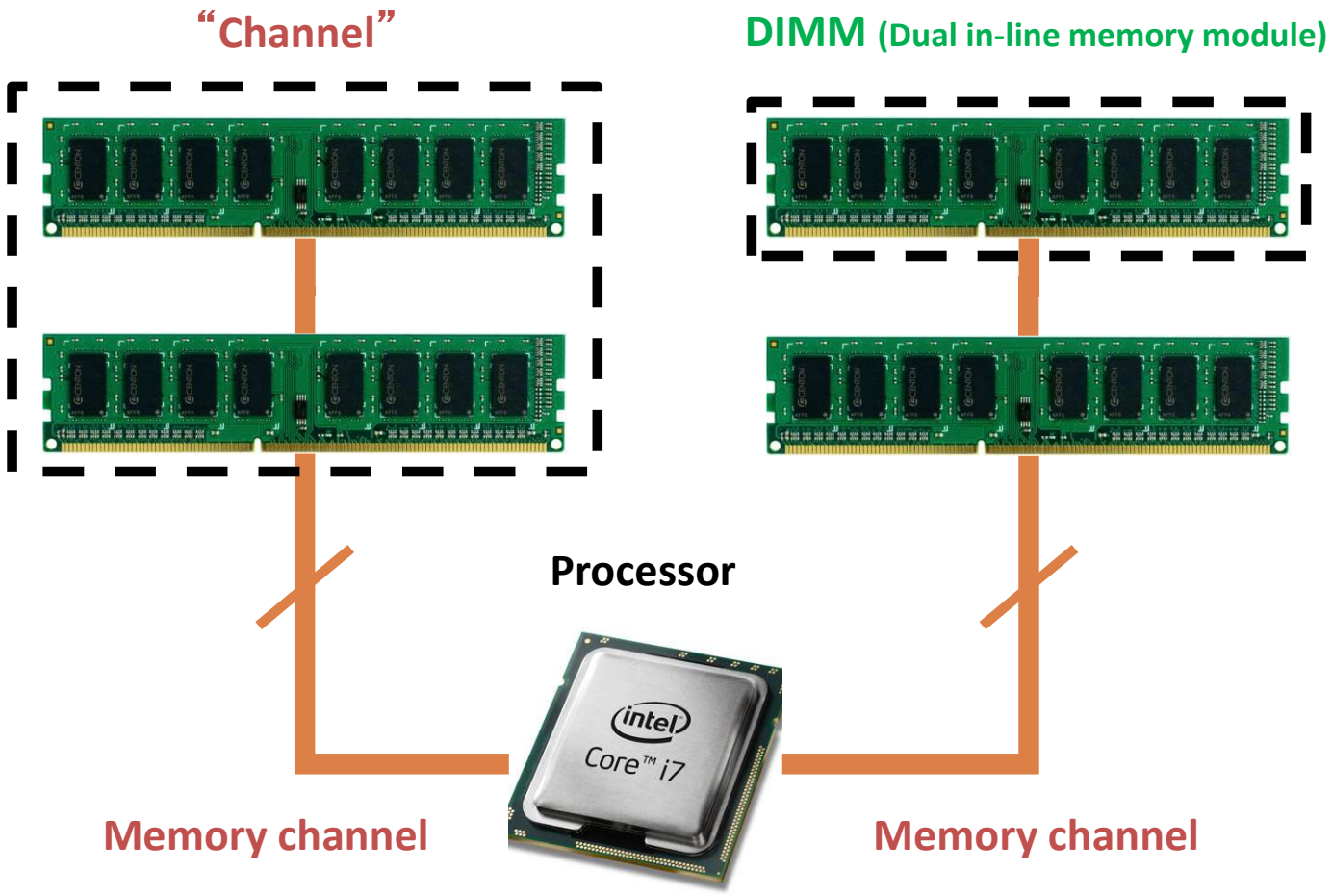


DRAM Subsystem Organization

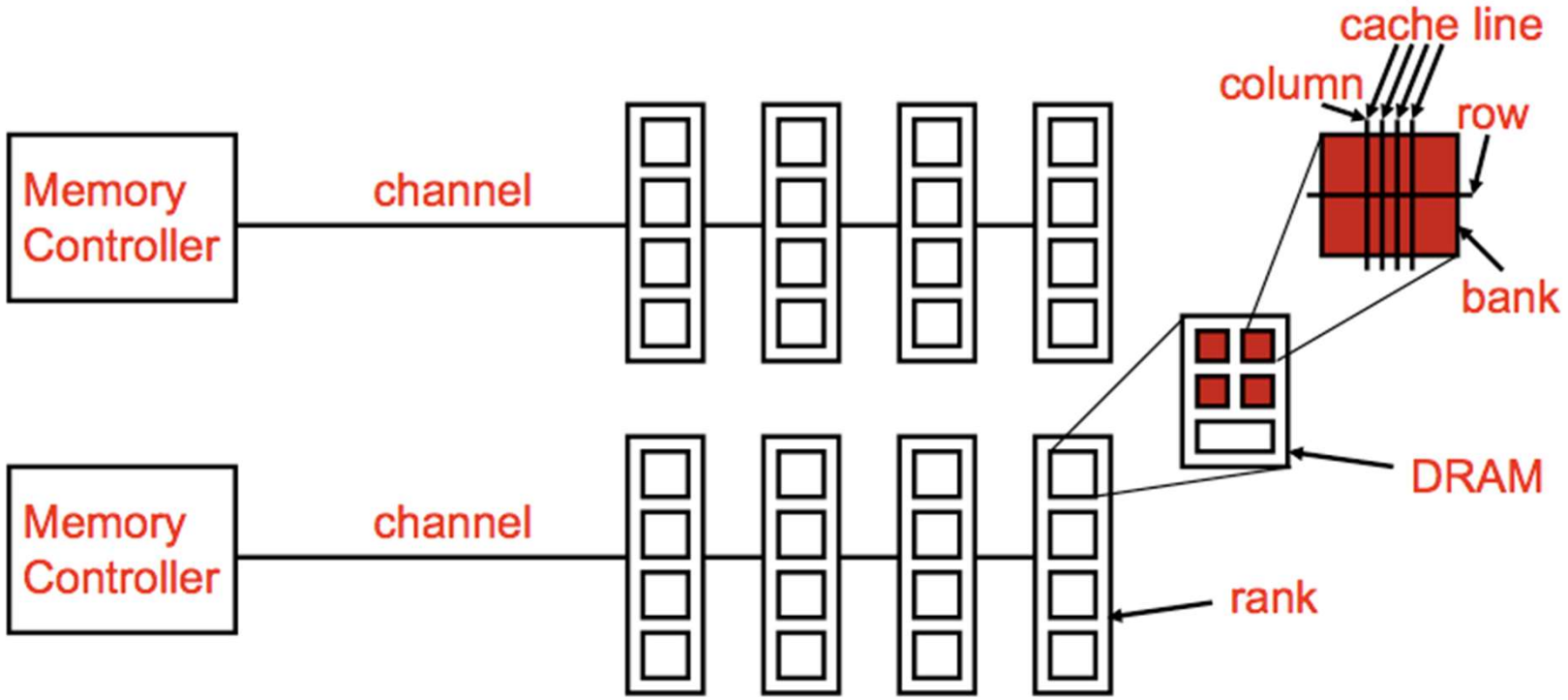
DRAM Subsystem Organization

- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/Column

The DRAM subsystem



Generalized Memory Structure

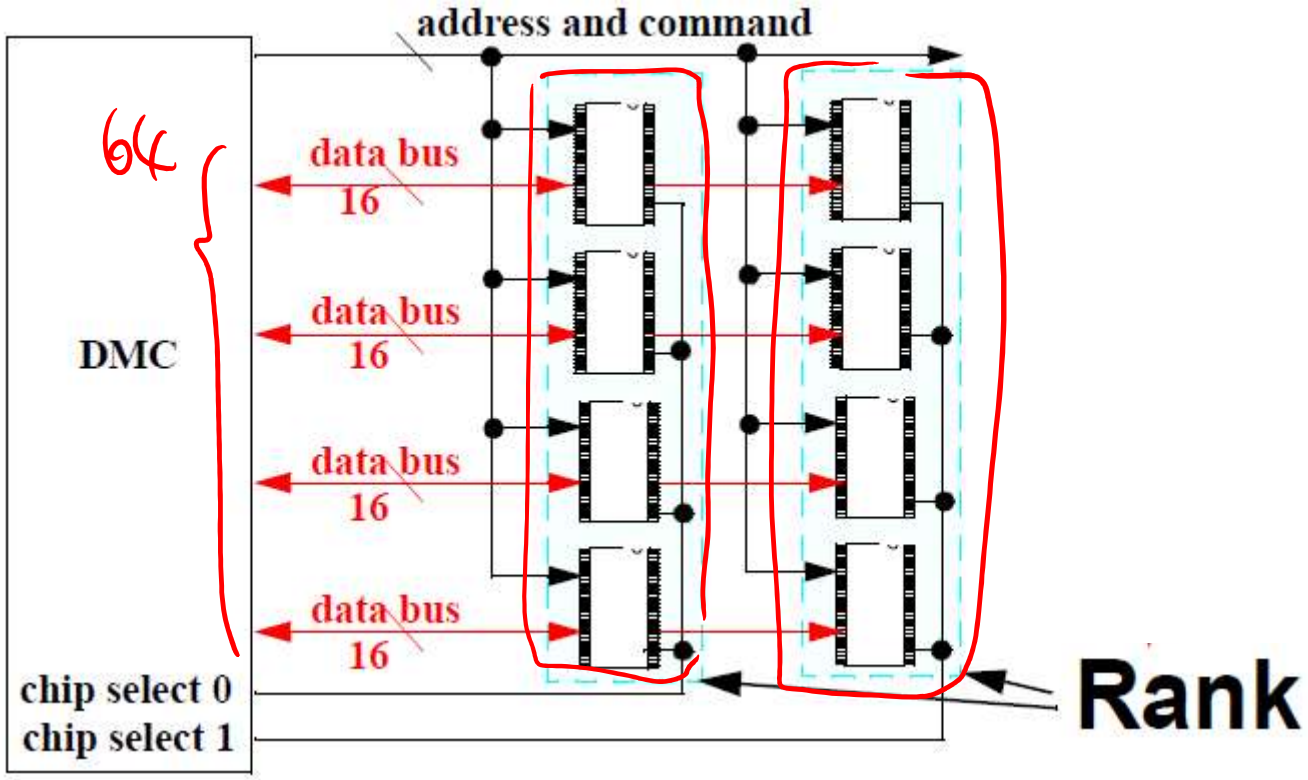


Rank and Module

- Rank: Multiple chips operated together to form a wide interface
- All chips comprising a rank are controlled at the same time
  - Respond to a single command
  - Share address and command buses, but provide different data
- A DRAM module consists of one or more ranks
  - E.g., DIMM (dual inline memory module)
  - This is what you plug into your motherboard
- If we have chips with 8-bit interface, to read 8 bytes in a single access, use 8 chips in a DIMM

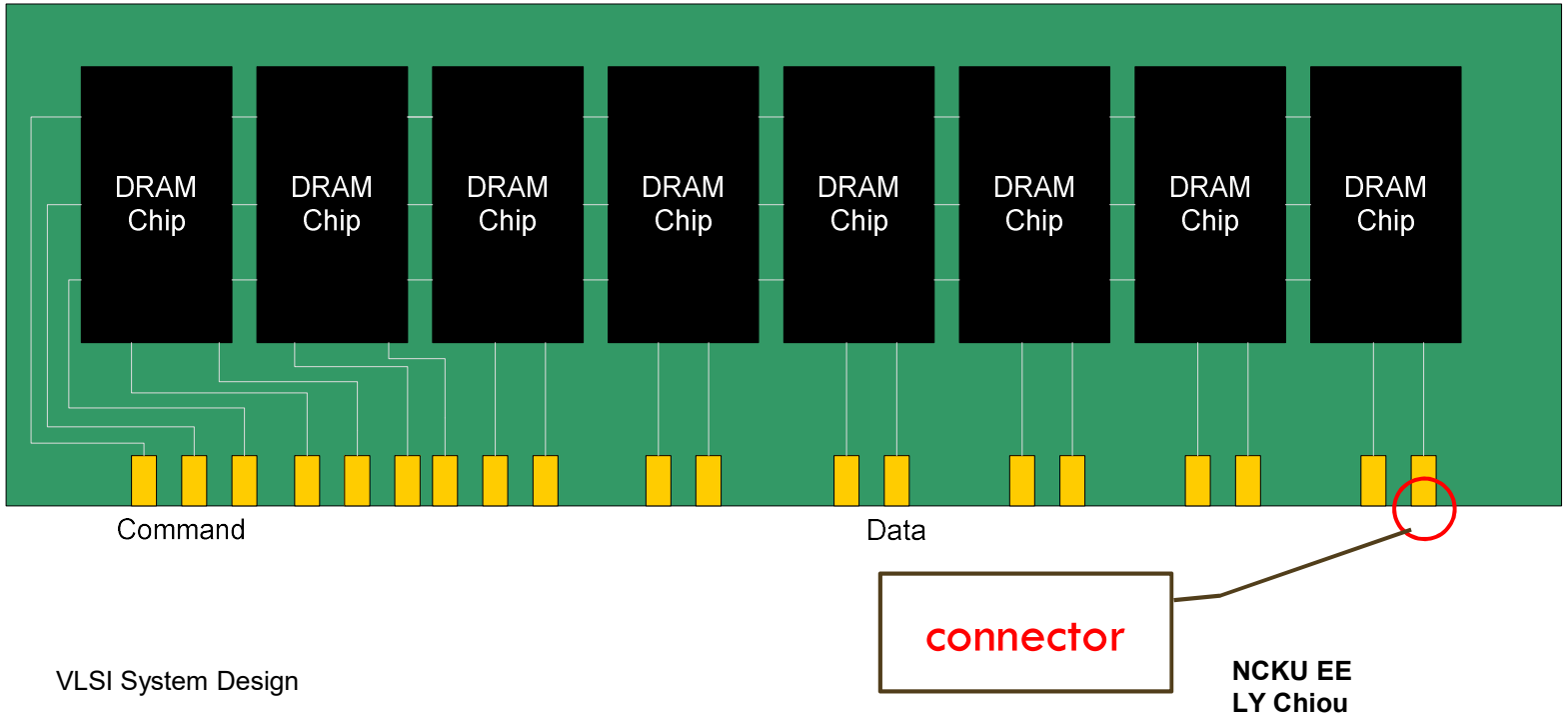
Rank

- This example has two ranks, four chips/rank, 16 bit data out/chip



A 64-bit Wide DIMM (One Rank)

- Dual-in-line memory module (DIMM), connectors on both are independent, i.e., not connected

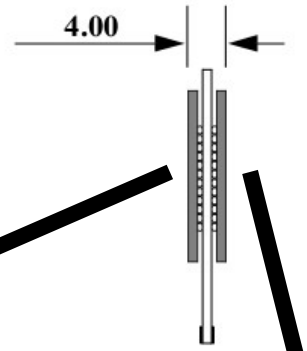


Breaking down a DIMM

DIMM (Dual in-line memory module)

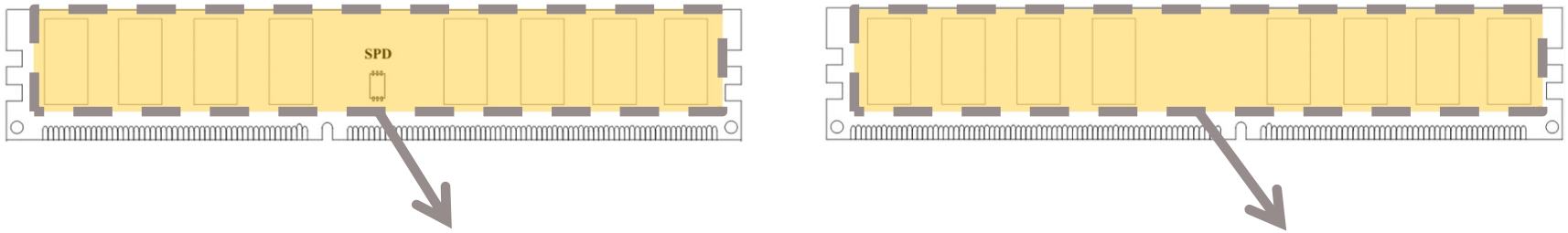
Breaking down a DIMM

DIMM (Dual in-line memory module)



Front of DIMM

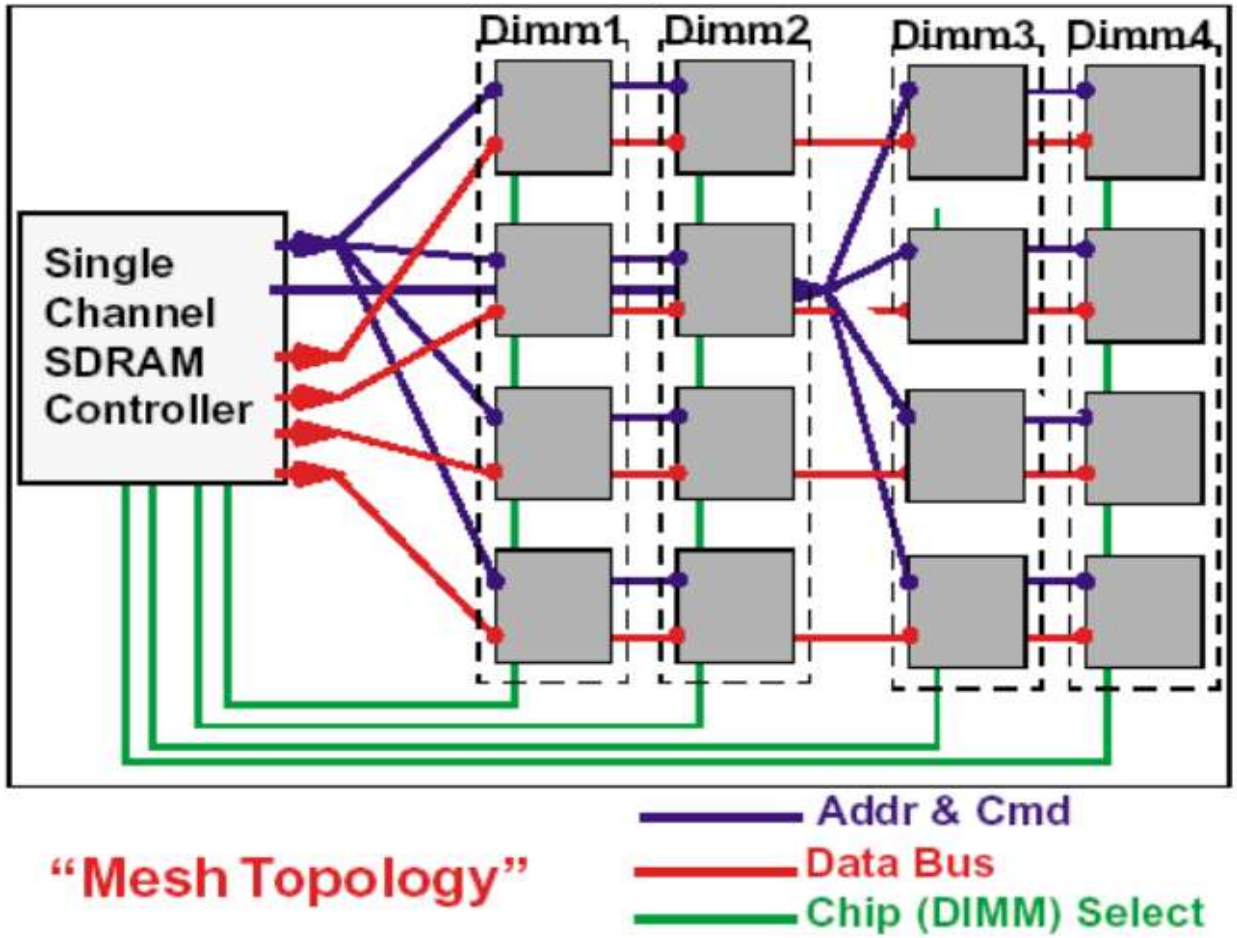
Back of DIMM



Rank 0: collection of 8 chips

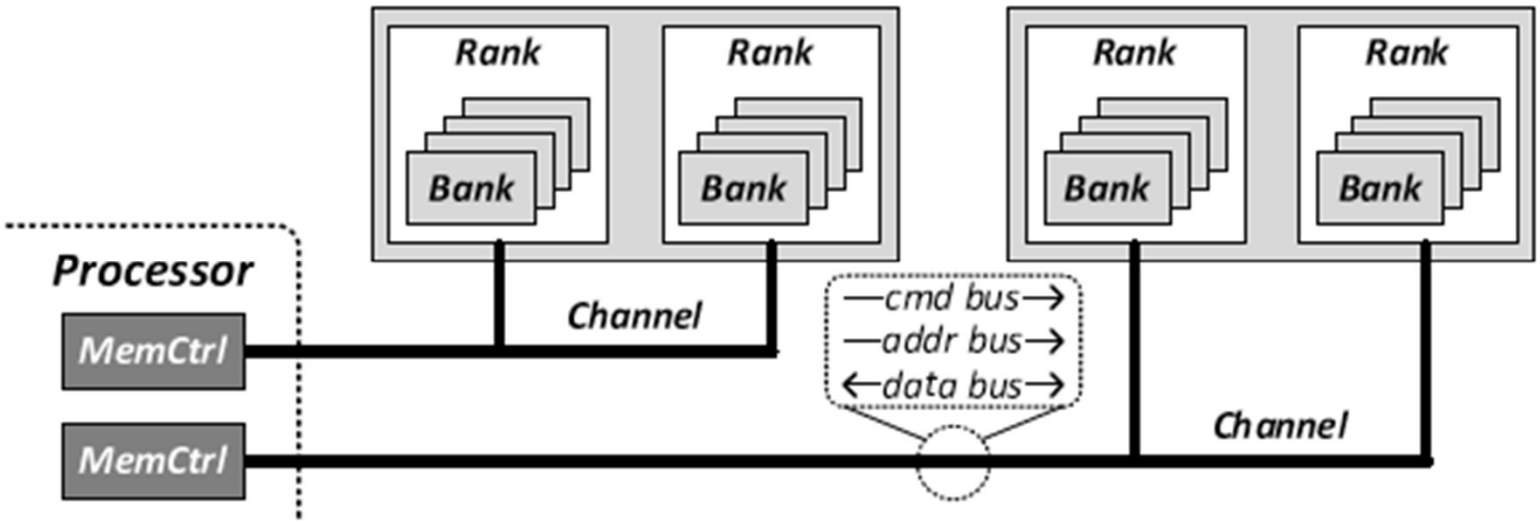
Rank 1

Multiple DIMMs

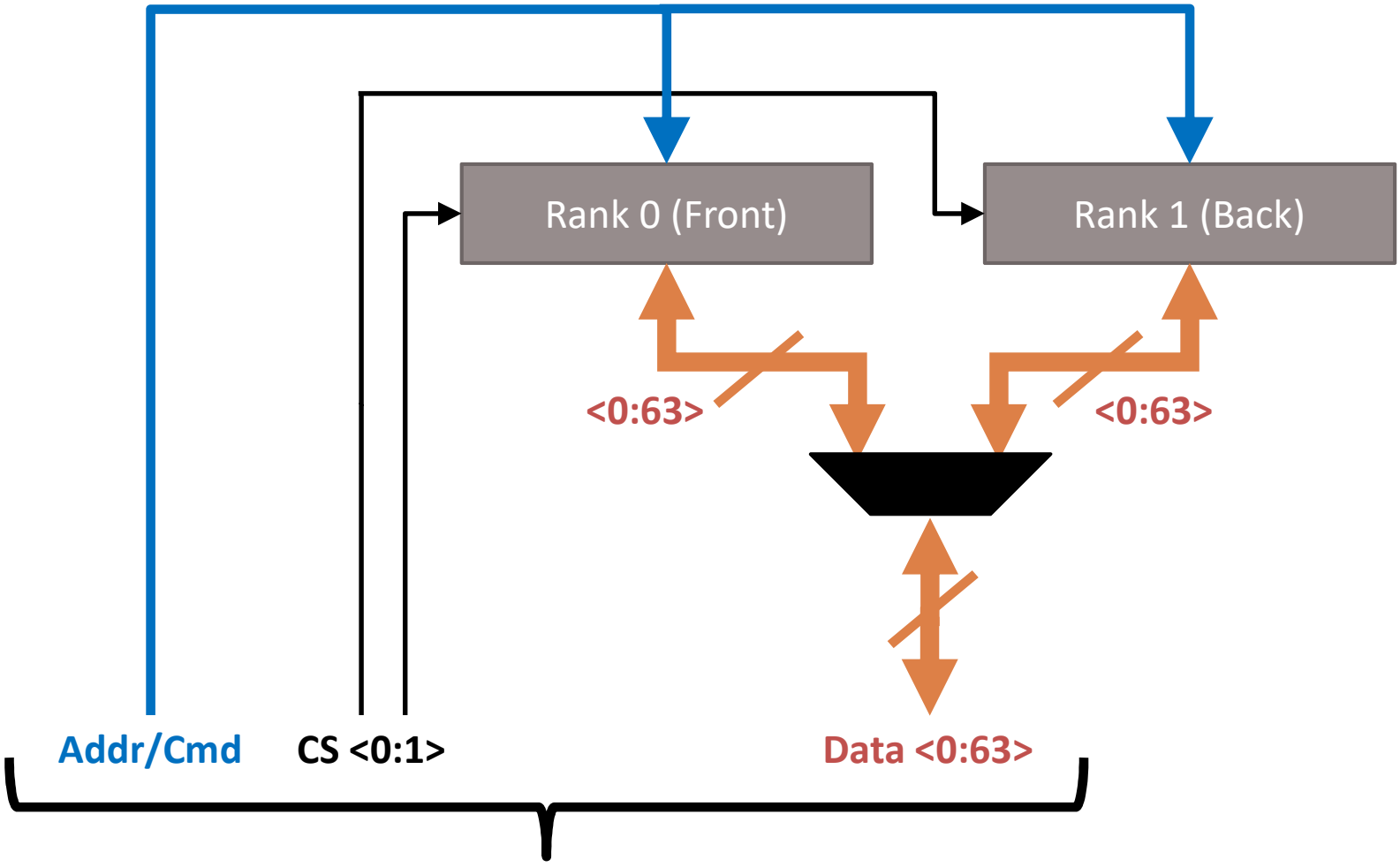


- Advantages:
  - Enables even higher capacity
- Disadvantages:
  - Interconnect complexity and energy consumption can be high

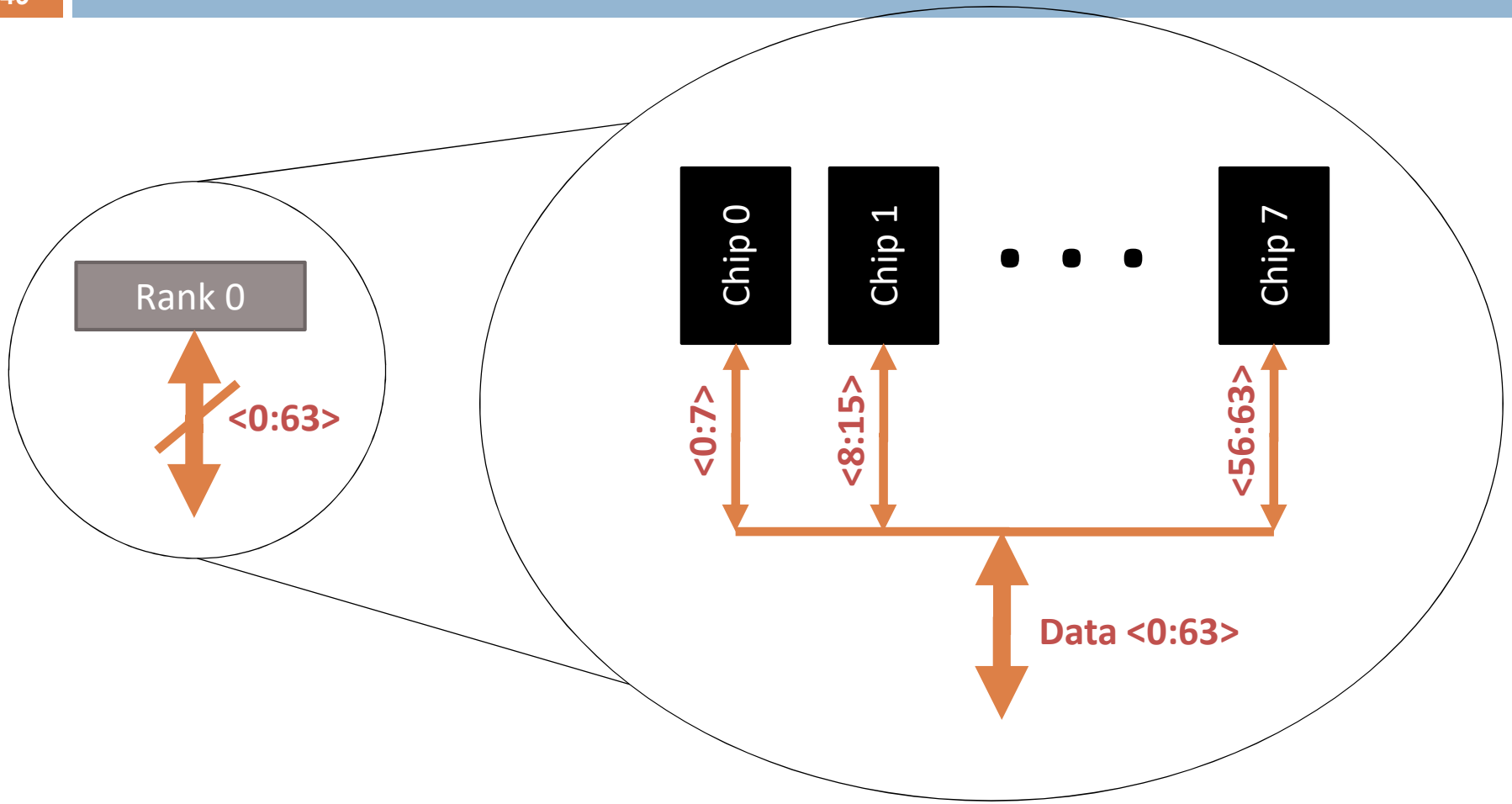
Generalized Memory Structure



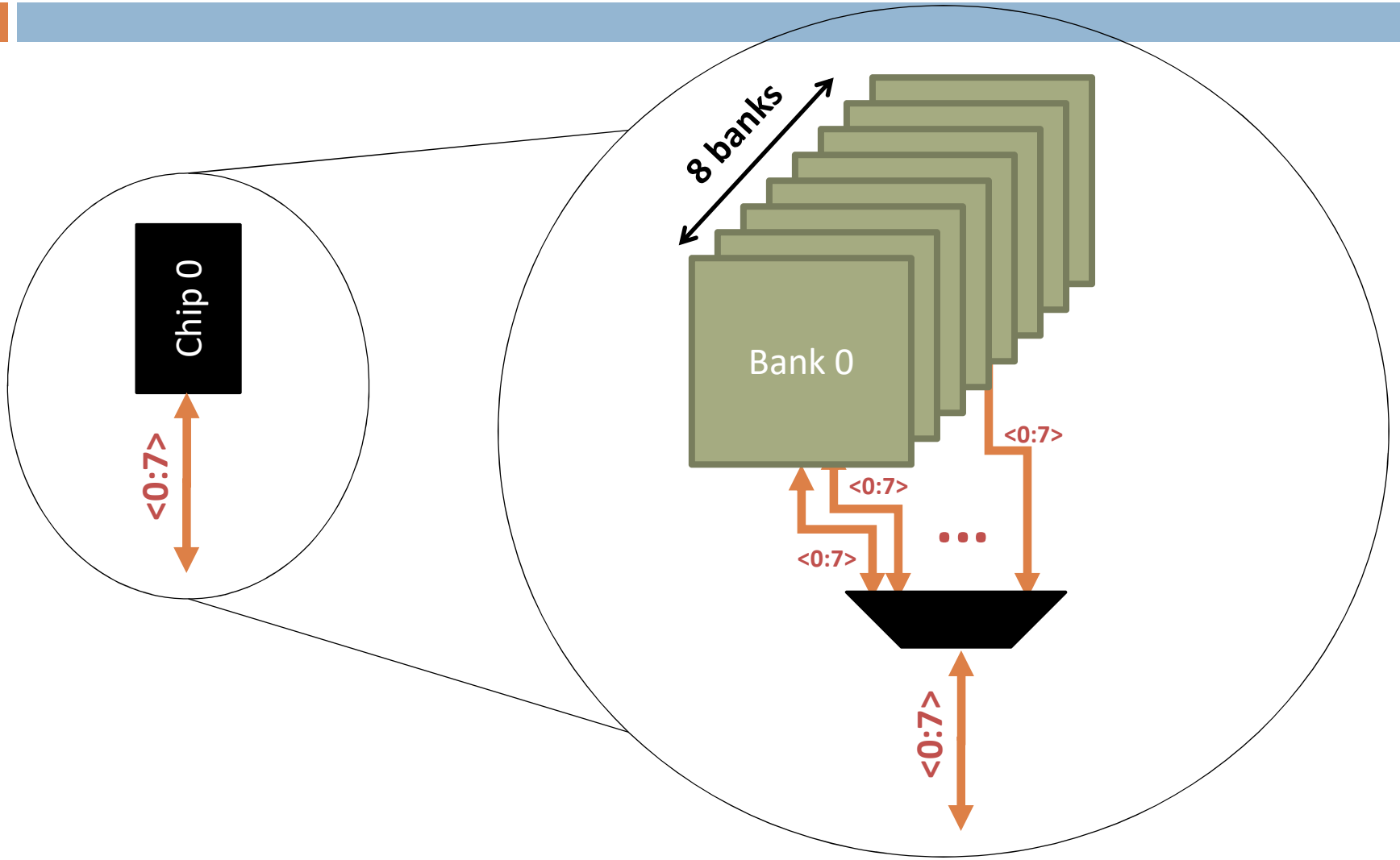
Rank


Rank 0 (Front)
Rank 1 (Back)
<0:63>
<0:63>
Addr/Cmd
CS <0:1>
Data <0:63>
Memory channel

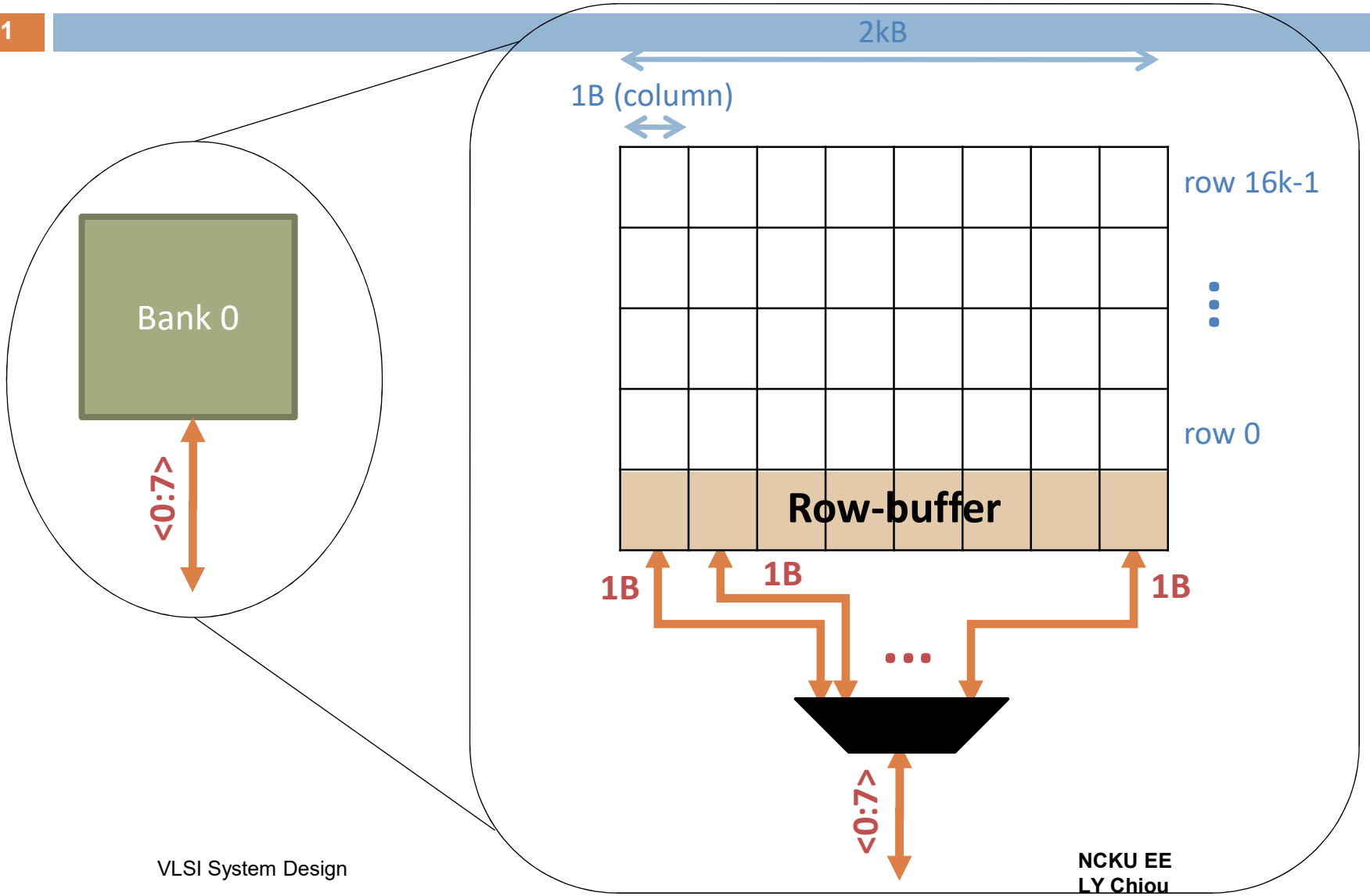
Breaking down a Rank



Breaking down a Chip

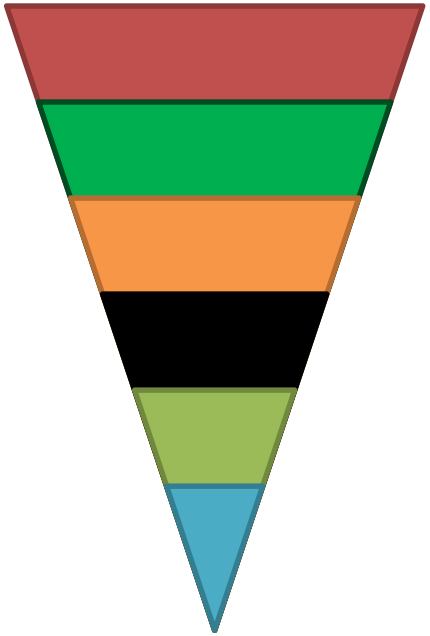


Breaking down a Bank

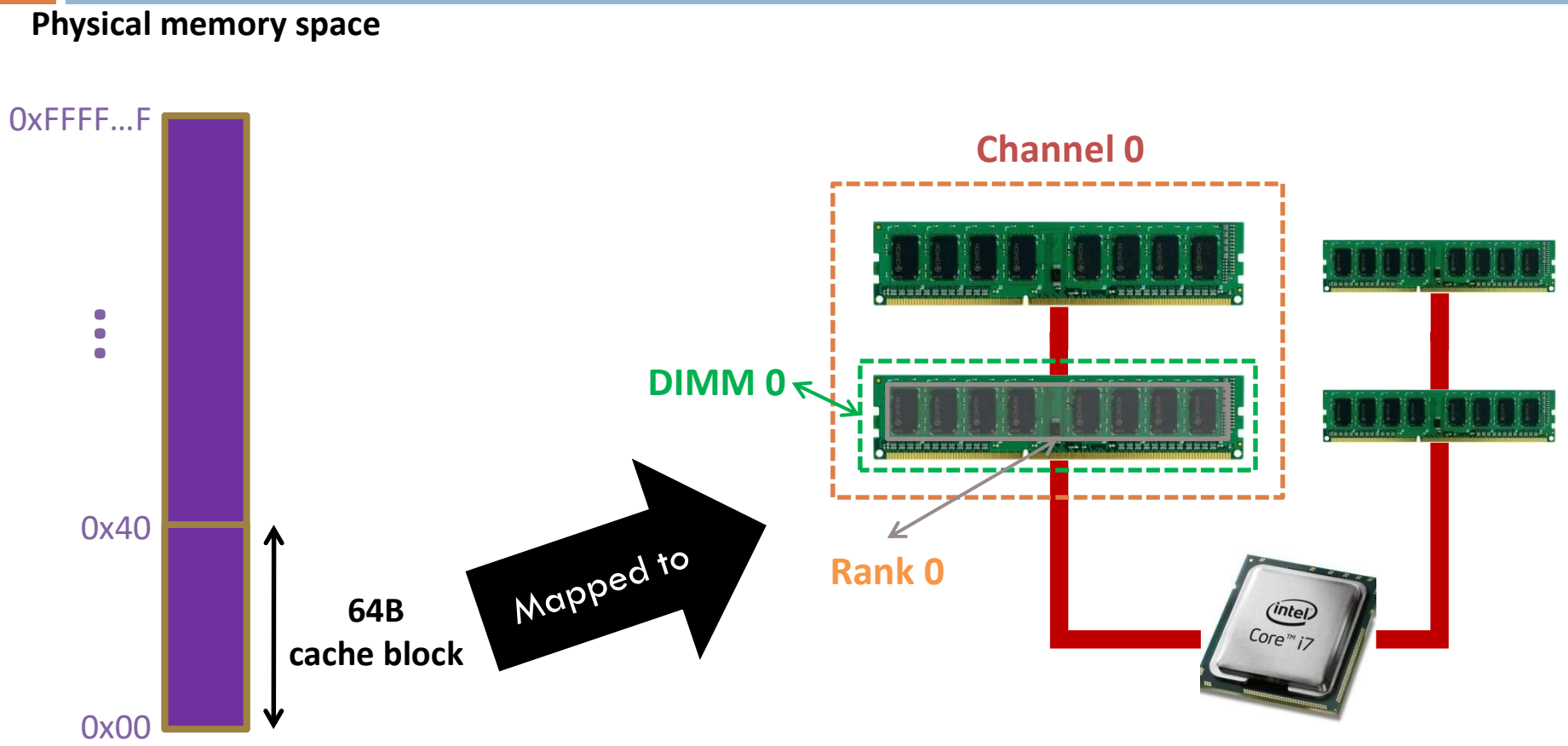


DRAM Subsystem Organization

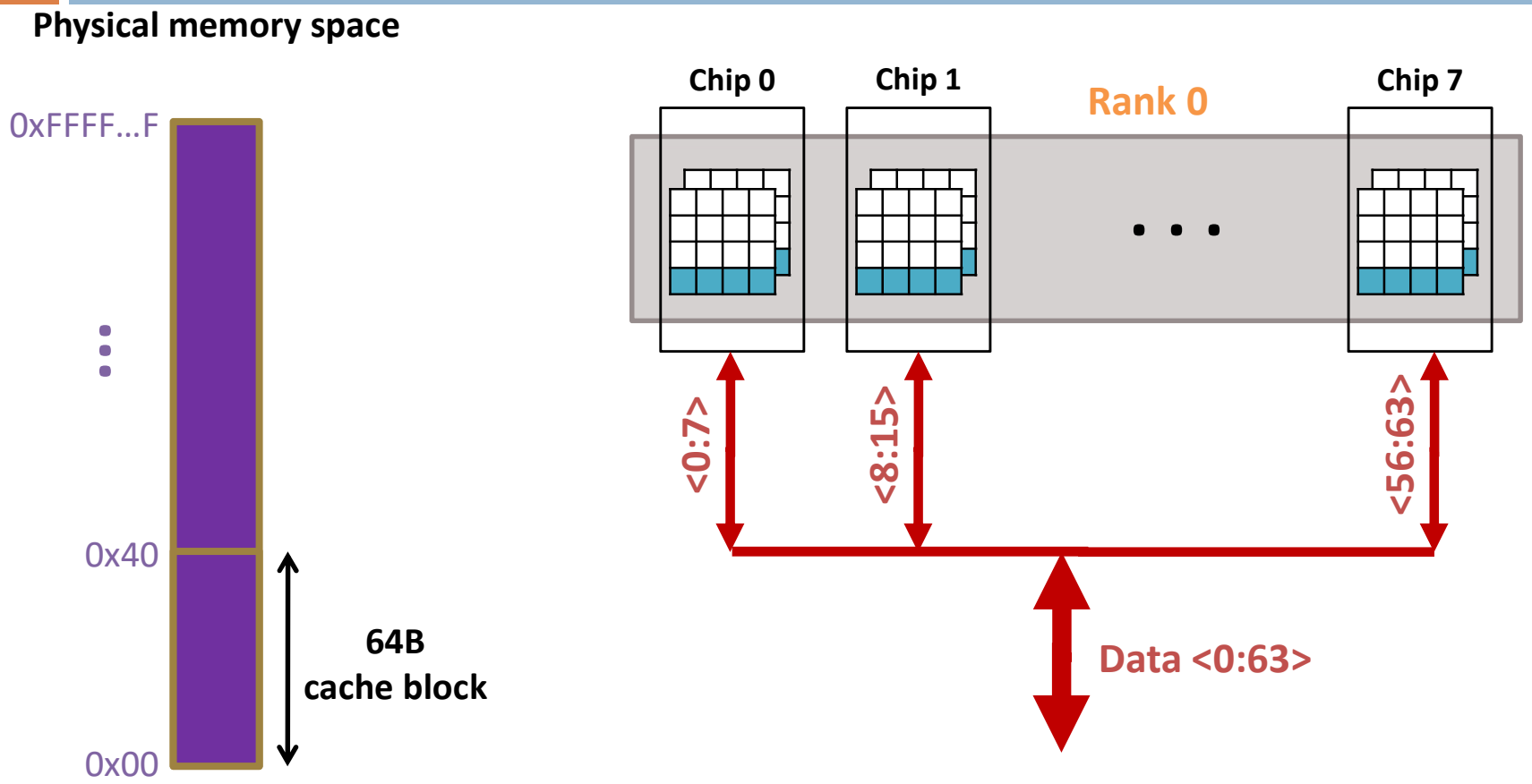
- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/Column



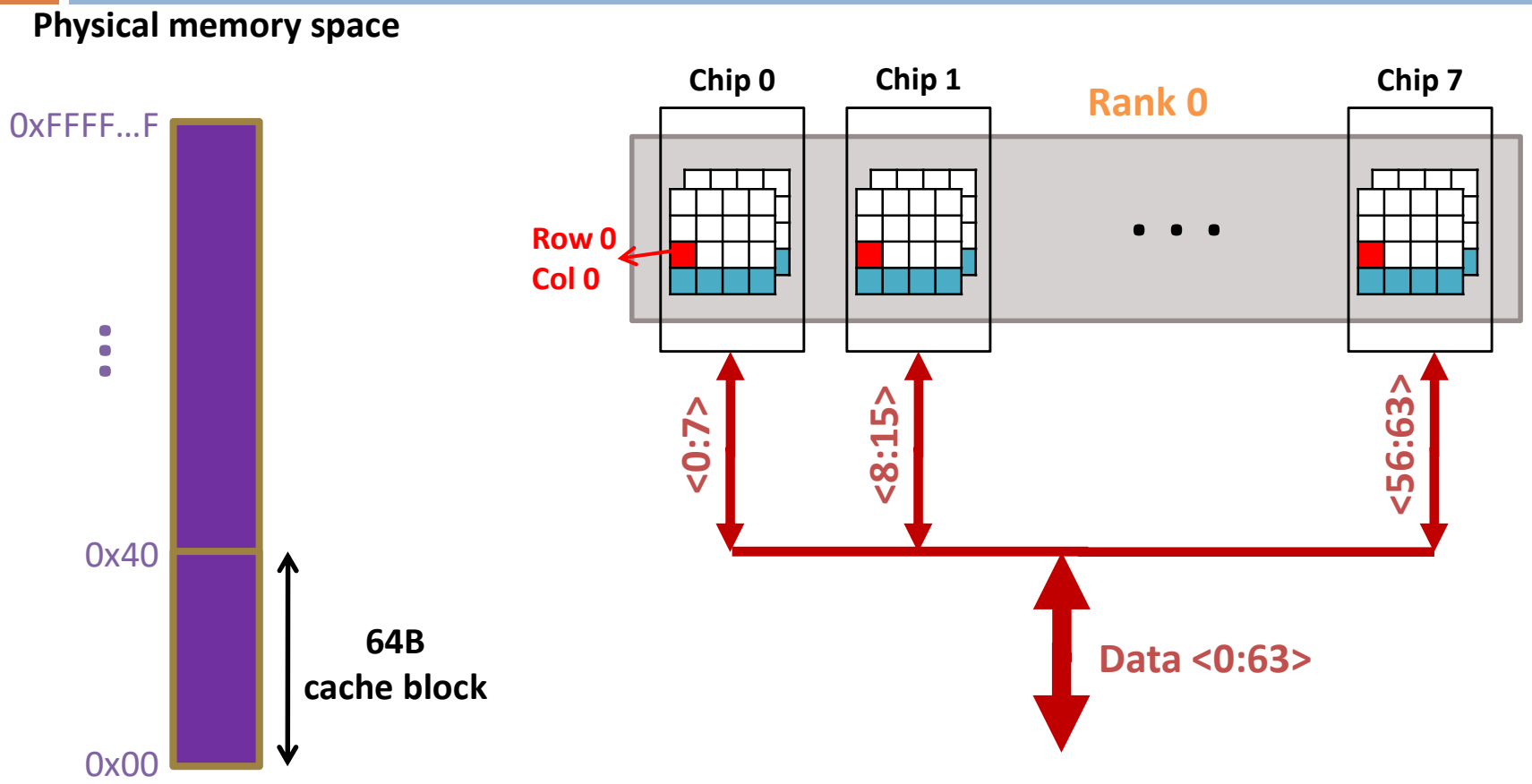
Example: Transferring a cache block



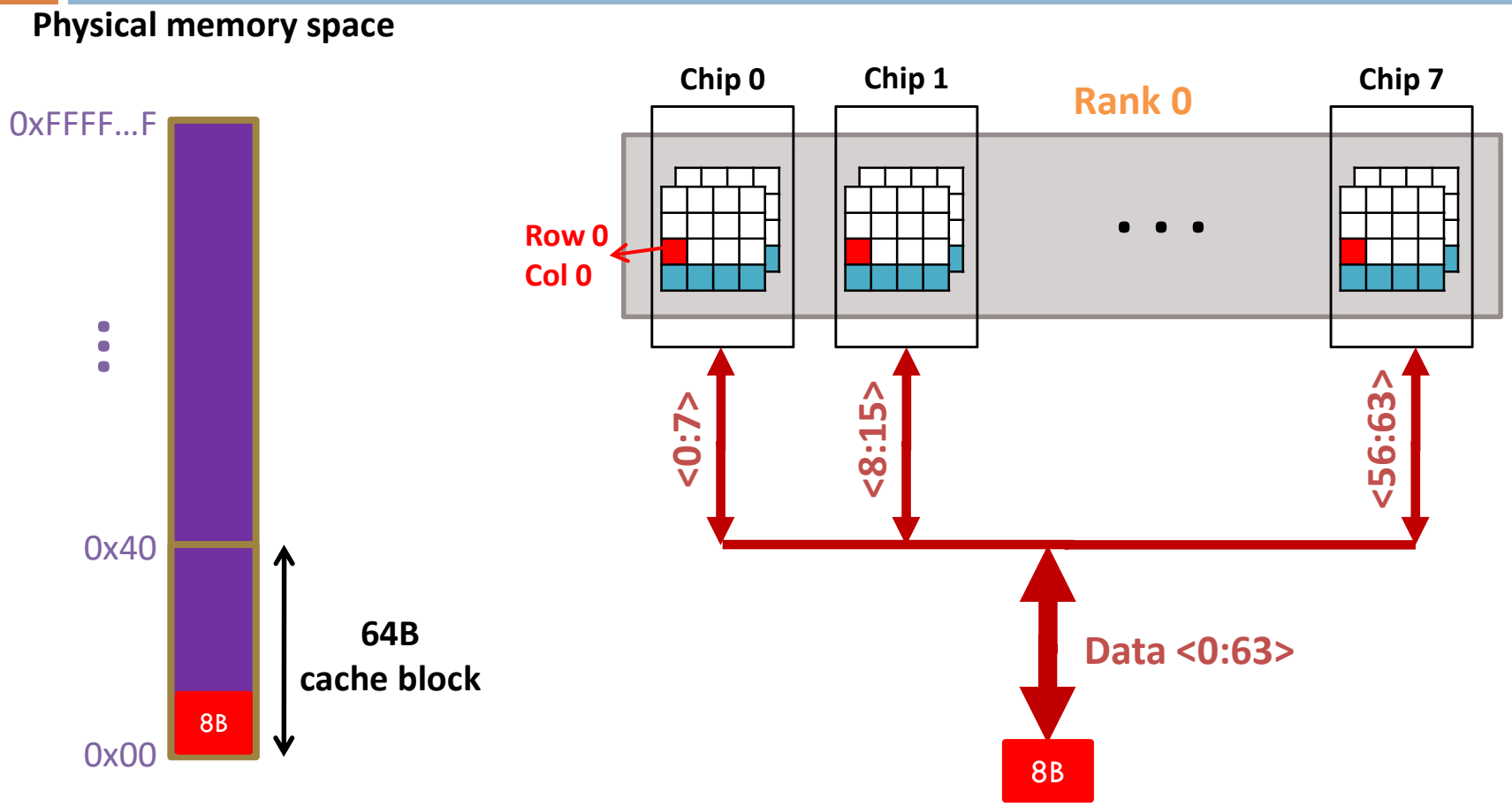
Example: Transferring a cache block



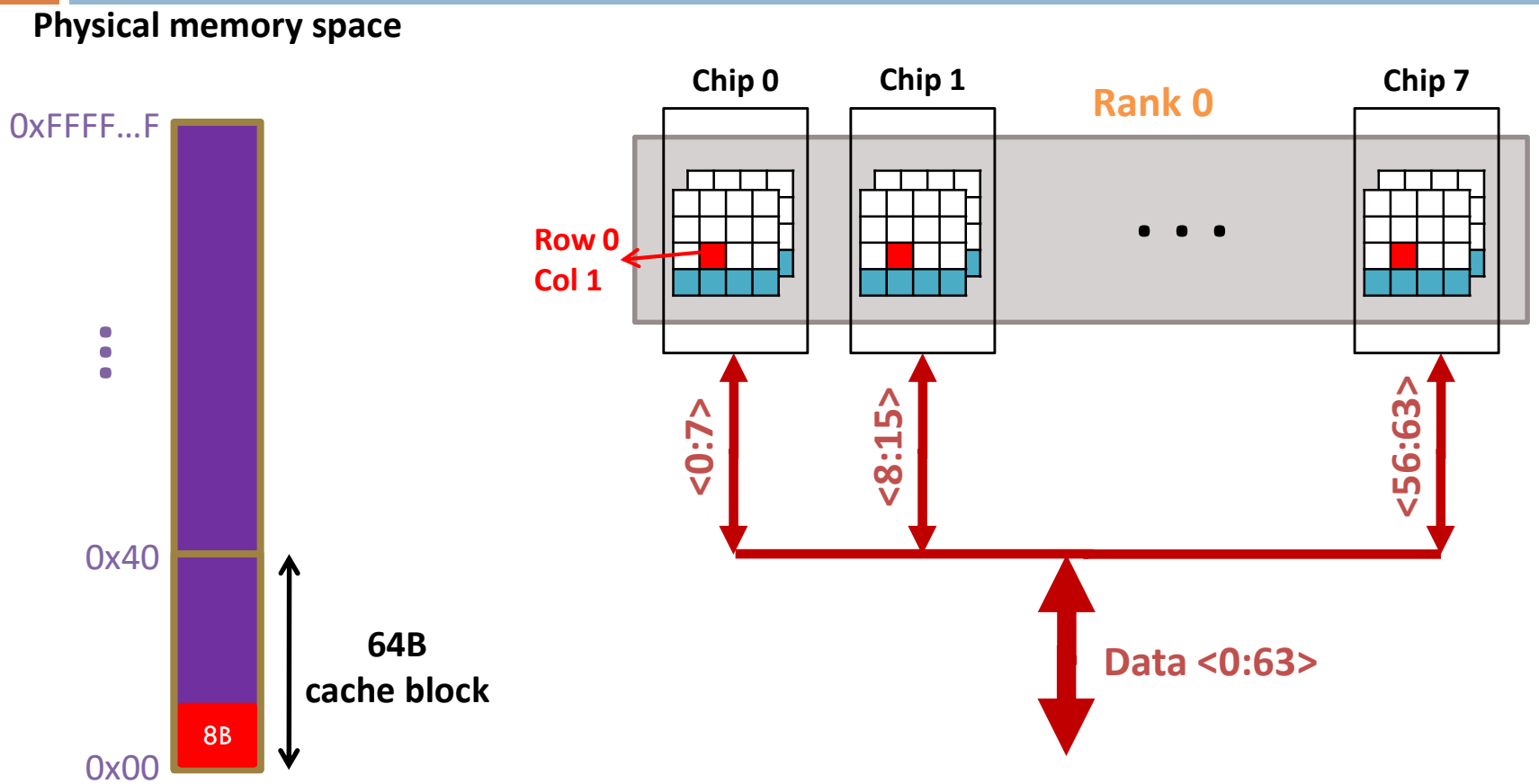
Example: Transferring a cache block



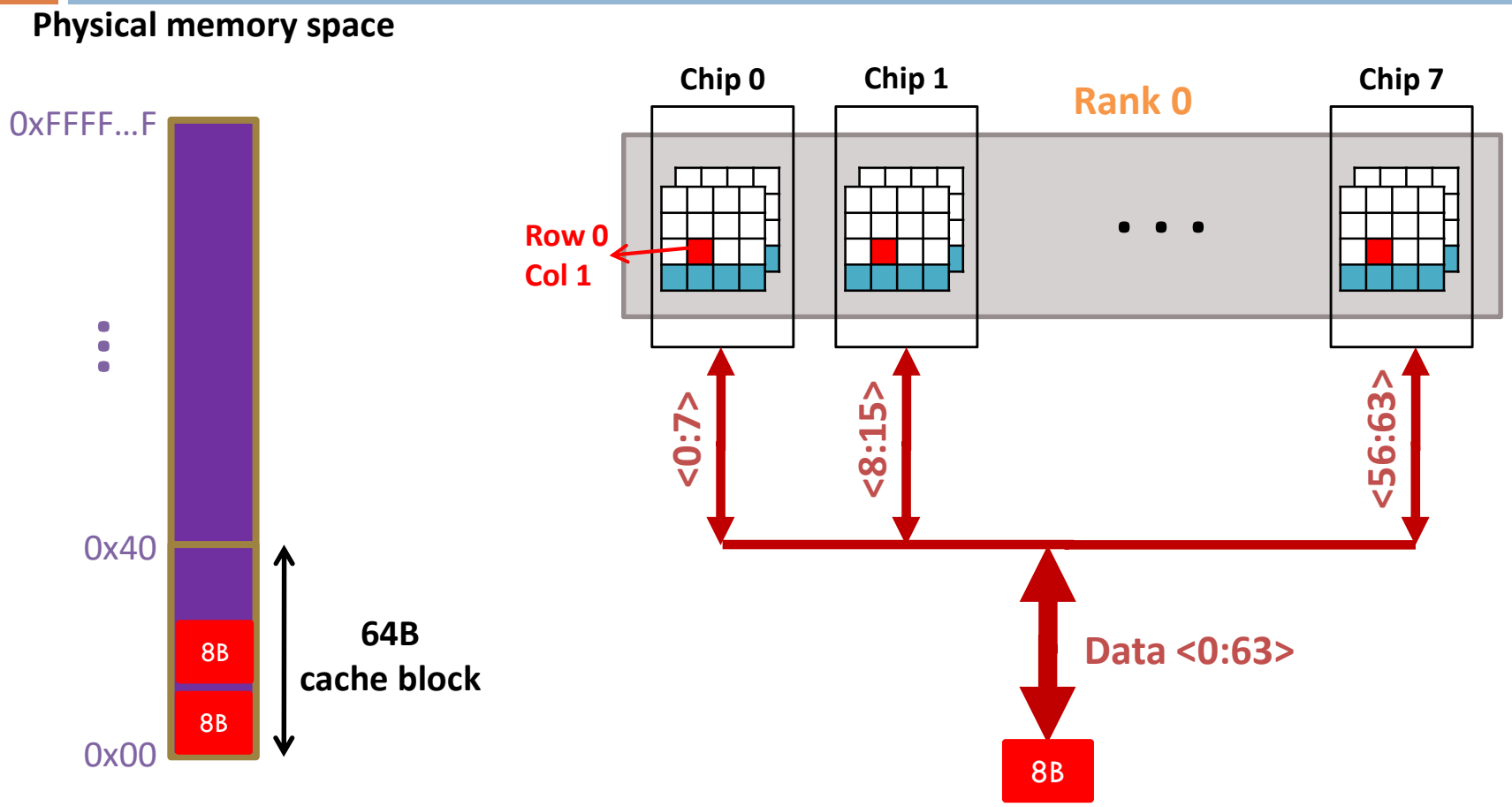
Example: Transferring a cache block



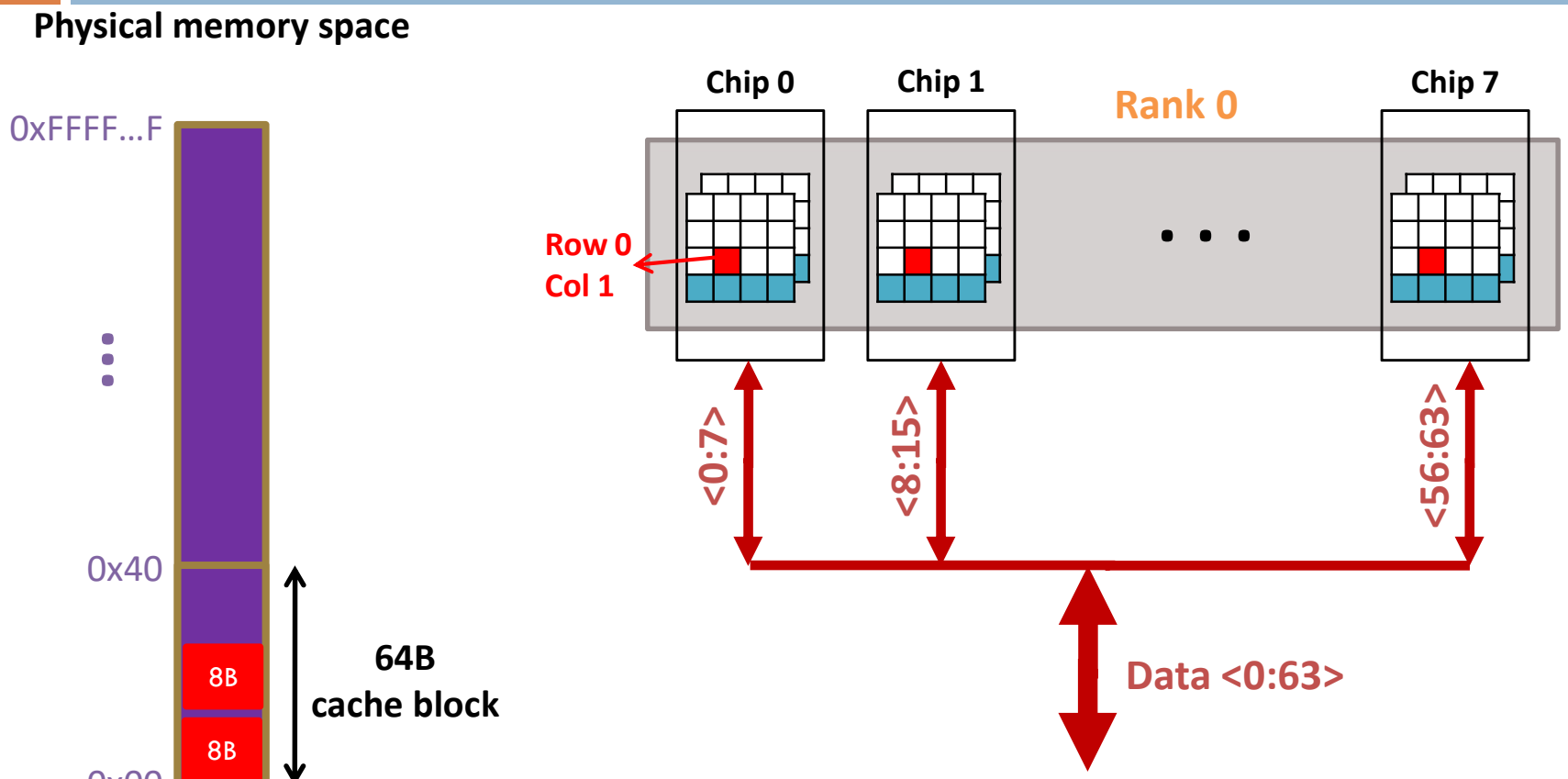
Example: Transferring a cache block



Example: Transferring a cache block



Example: Transferring a cache block



A 64B cache block takes 8 I/O cycles to transfer.

During the process, 8 columns are read sequentially.

Latency Components: Basic DRAM Operation

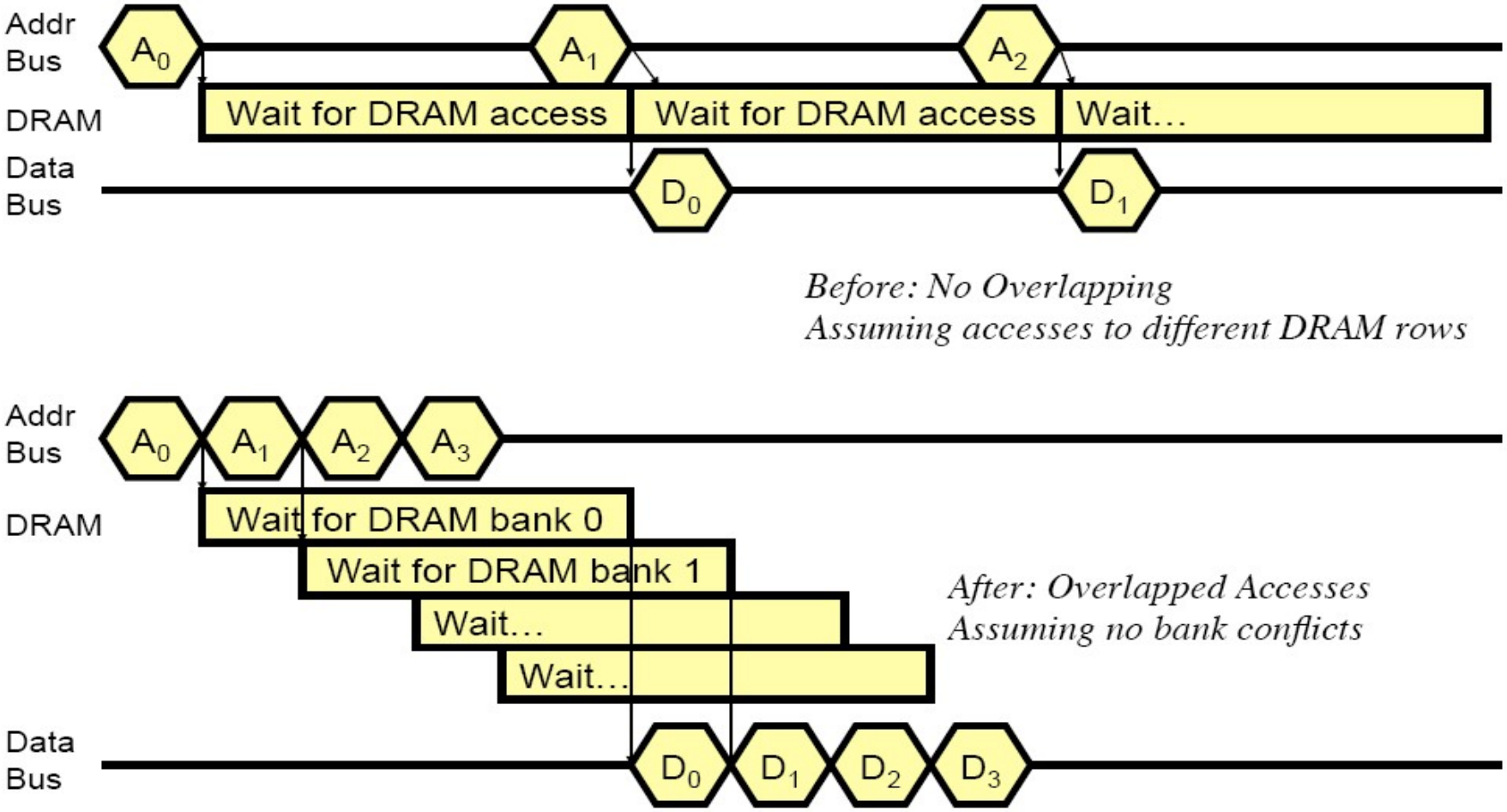
- CPU → controller transfer time
- Controller latency
  - Queuing & scheduling delay at the controller
  - Access converted to basic commands
- Controller → DRAM transfer time
- DRAM bank latency
  - Simple CAS if row is "open" OR
  - RAS + CAS if array precharged OR
  - PRE + RAS + CAS (worst case)
- DRAM → CPU transfer time (through controller)

Multiple Banks (Interleaving) and Channels

- Multiple banks
  - Enable concurrent DRAM accesses
  - Bits in address determine which bank an address resides in
- Multiple independent channels serve the same purpose
  - But they are even better because they have separate data buses
  - Increased bus bandwidth

- Enabling more concurrency requires reducing
  - Bank conflicts
  - Channel conflicts
- How to select/randomize bank/channel indices in address?
  - Lower order bits have more entropy
  - Randomizing hash functions (XOR of different address bits)

How Multiple Banks/Channels Help



Multiple Channels

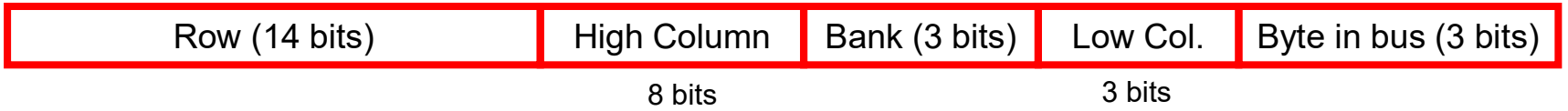
- Advantages
  - Increased bandwidth
  - Multiple concurrent accesses (if independent channels)

- Disadvantages
  - Higher cost than a single channel
    - More board wires
    - More pins (if on-chip memory controller)

Address Mapping (Single Channel)

- Single-channel system with 8-byte memory bus
  - 2GB memory, 8 banks, 16K rows & 2K columns per bank
- Row interleaving
  - Consecutive rows of memory in consecutive banks

- Cache block interleaving
  - Consecutive cache block addresses in consecutive banks
  - 64 byte cache blocks

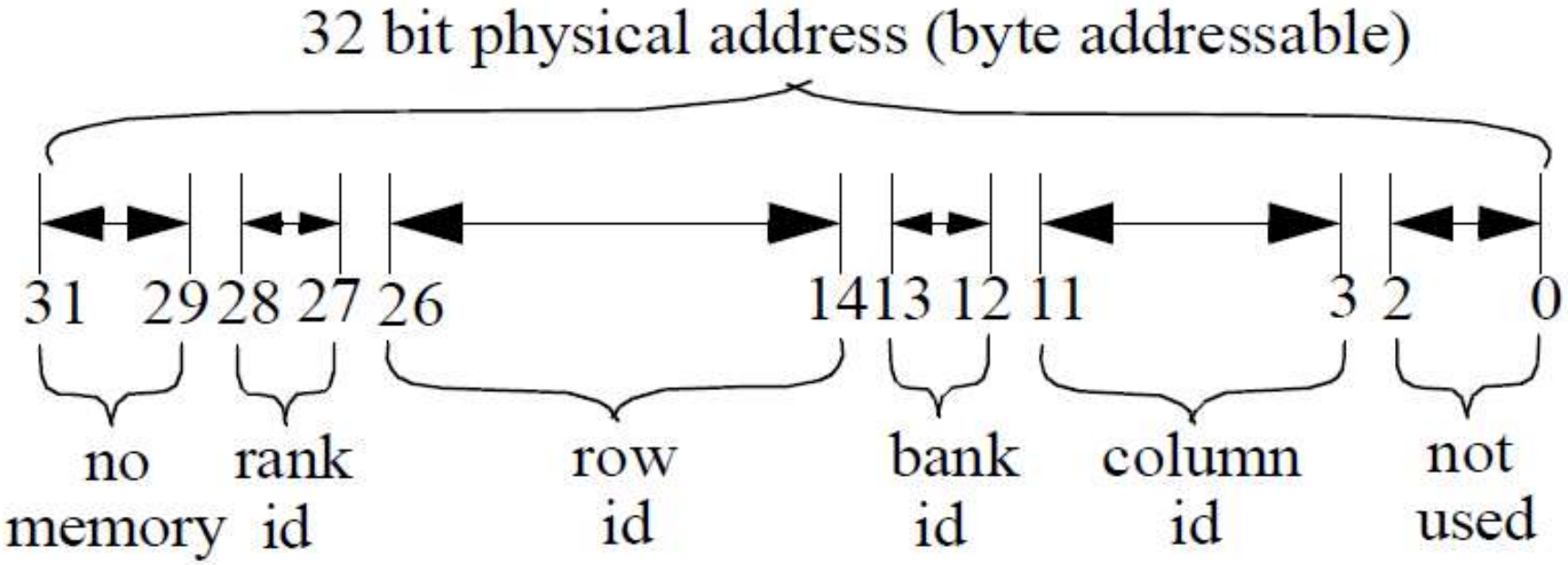


  - Accesses to consecutive cache blocks can be serviced in parallel
  - How about random accesses? Strided accesses?

DRAM Basics

- Address multiplexing
  - Send row address when RAS asserted
  - Send column address when CAS asserted
- DRAM reads are self-destructive
  - Rewrite after a read
- Memory array
  - All bits within an array work in unison
- Memory bank
  - Different banks can operate independently
- DRAM rank
  - Chips inside the same rank are accessed simultaneously

DRAM Address Structure



Notes

- The memory controller schedules memory accesses to maximize row buffer hit rates and bank/rank parallelism
- Banks and ranks offer memory parallelism
- Row buffers act as a cache within DRAM
  - Row buffer hit: ~20 ns access time (must only move data from row buffer to pins)
  - Empty row buffer access: ~40 ns (must first read arrays, then move data from row buffer to pins)
  - Row buffer conflict: ~60 ns (must first writeback the existing row, then read new row, then move data to pins)
- In addition, must wait in the queue (tens of nano-seconds) and incur address/cmd/data transfer delays (~10 ns)

Latency and Power Wall

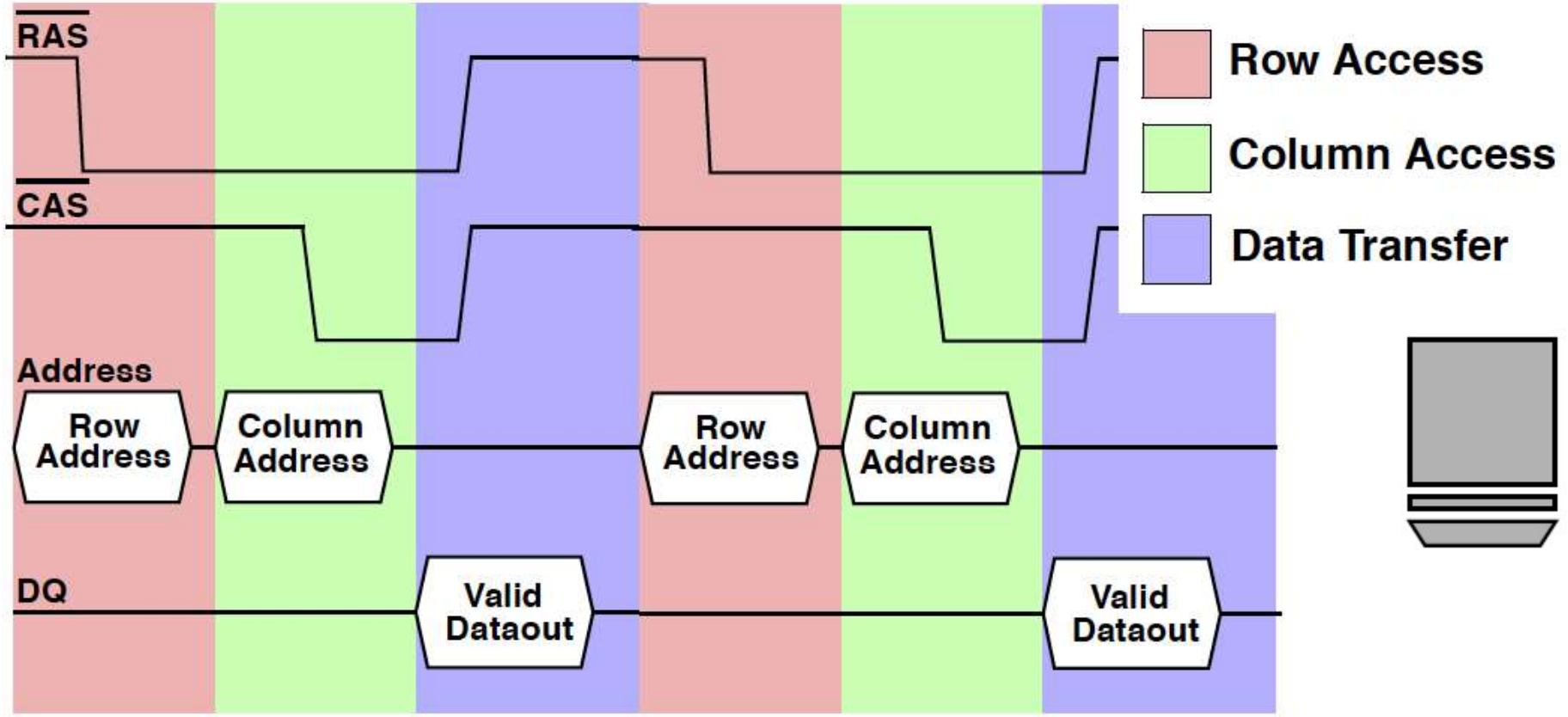
- Both improved by employing smaller arrays
  - Penalty in density and cost
- Both improved by increasing the row buffer hit rate
  - Requires intelligent mapping of data to rows, clever scheduling of requests, etc.

DRAM System Signaling and Timing

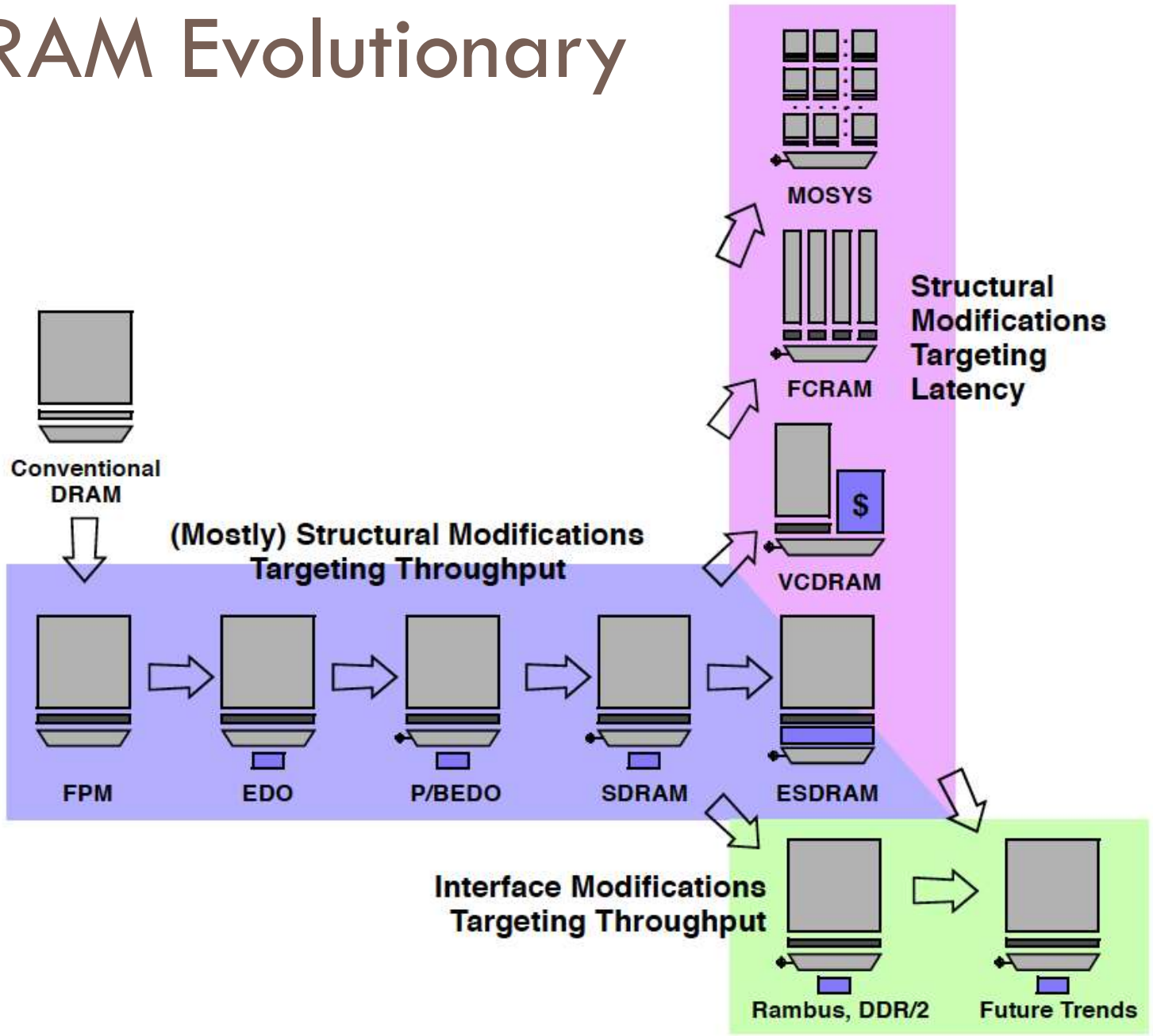
Different DRAM Systems

- Innovation targeted towards higher bandwidth for memory systems:
  - SDRAM - synchronous DRAM
  - RDRAM - Rambus DRAM
  - EDORAM - extended data out SRAM
  - Three-dimensional RAM
  - Hyper-page mode DRAM video RAM
  - Multibank DRAM

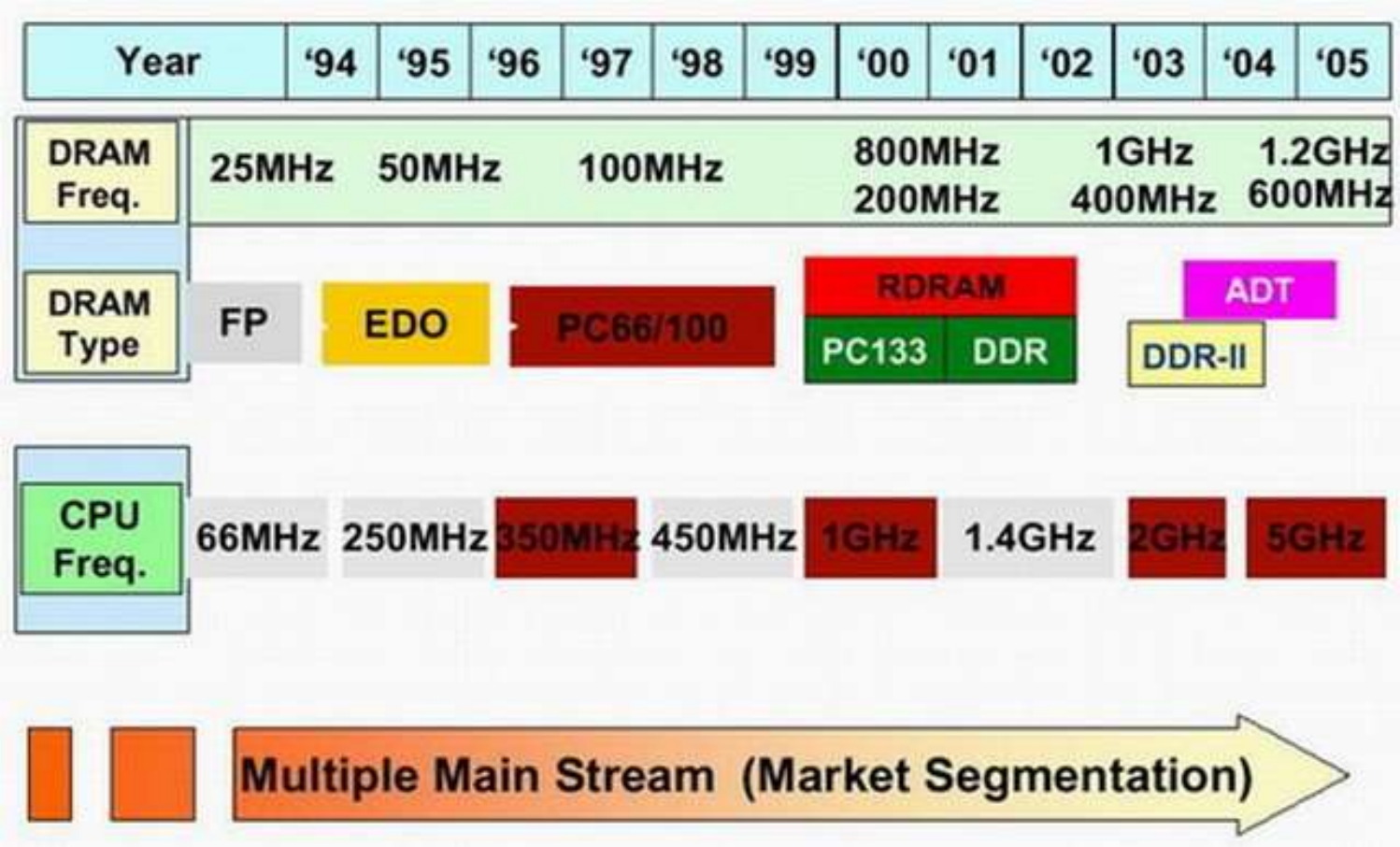
Read Timing of Conventional DRAM



DRAM Evolutionary

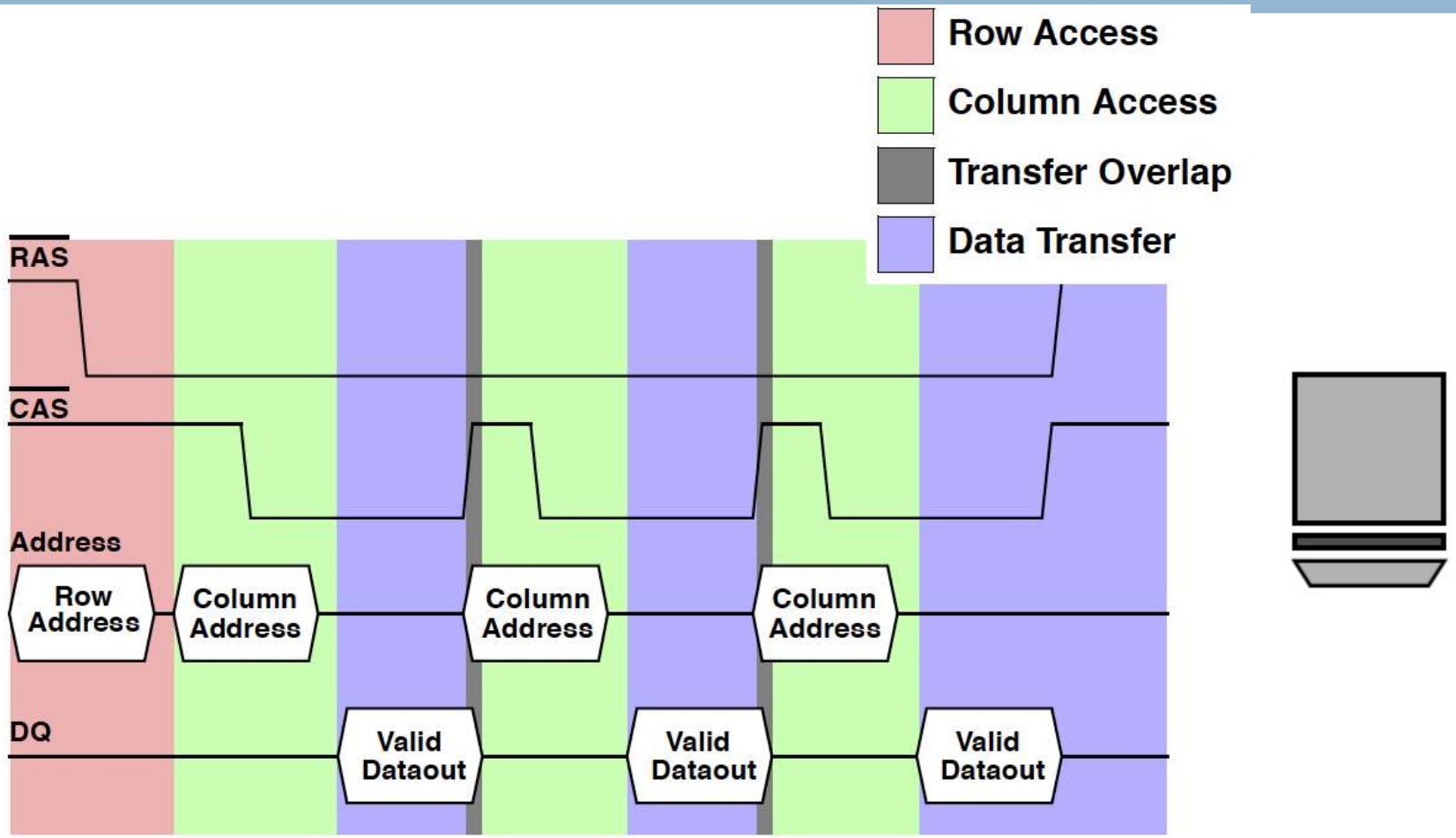


Frequency of DRAM Generations

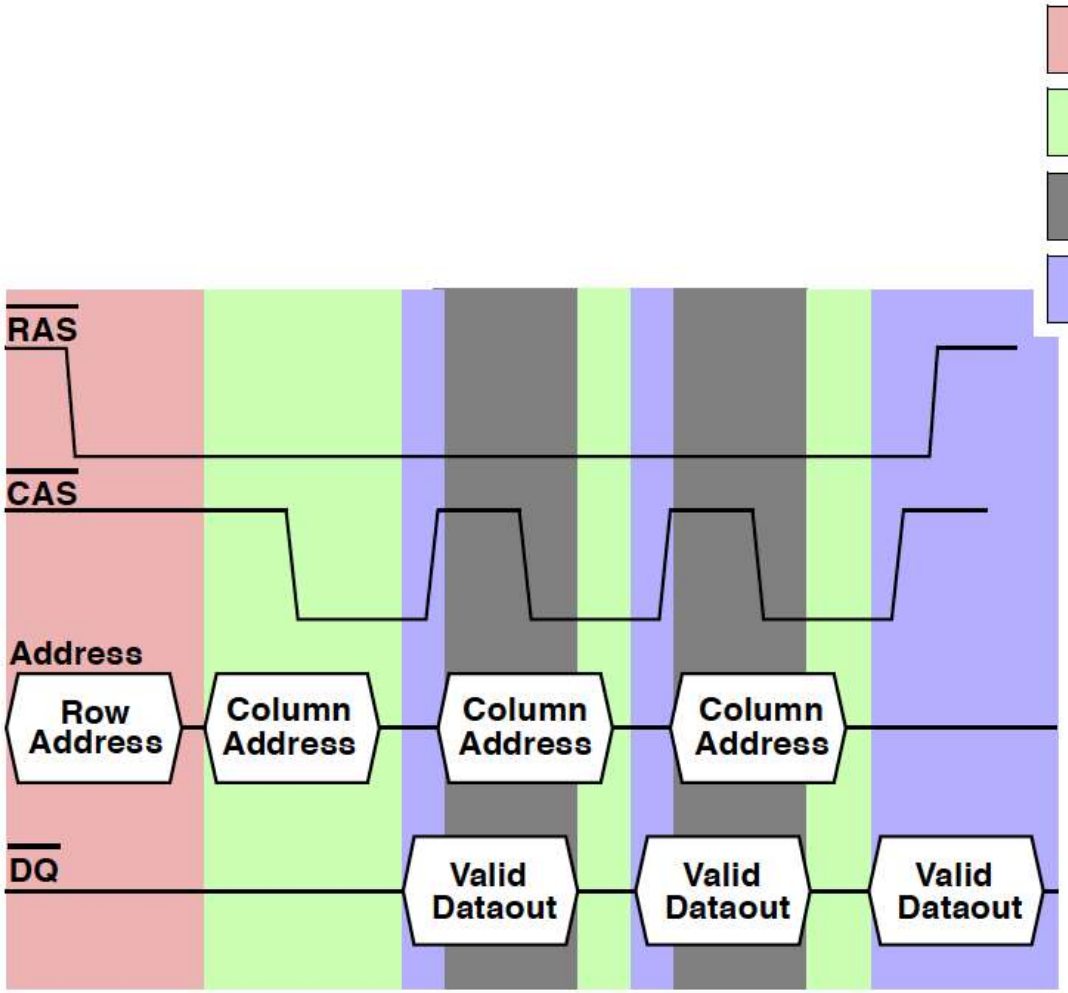


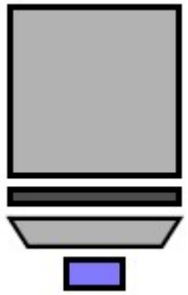
Source: "記憶體10年技術演進史，系統顆粒DDR與顯示顆粒GDDR差在哪？" by Tandee on internet

Reading Timing for Fast Page Mode



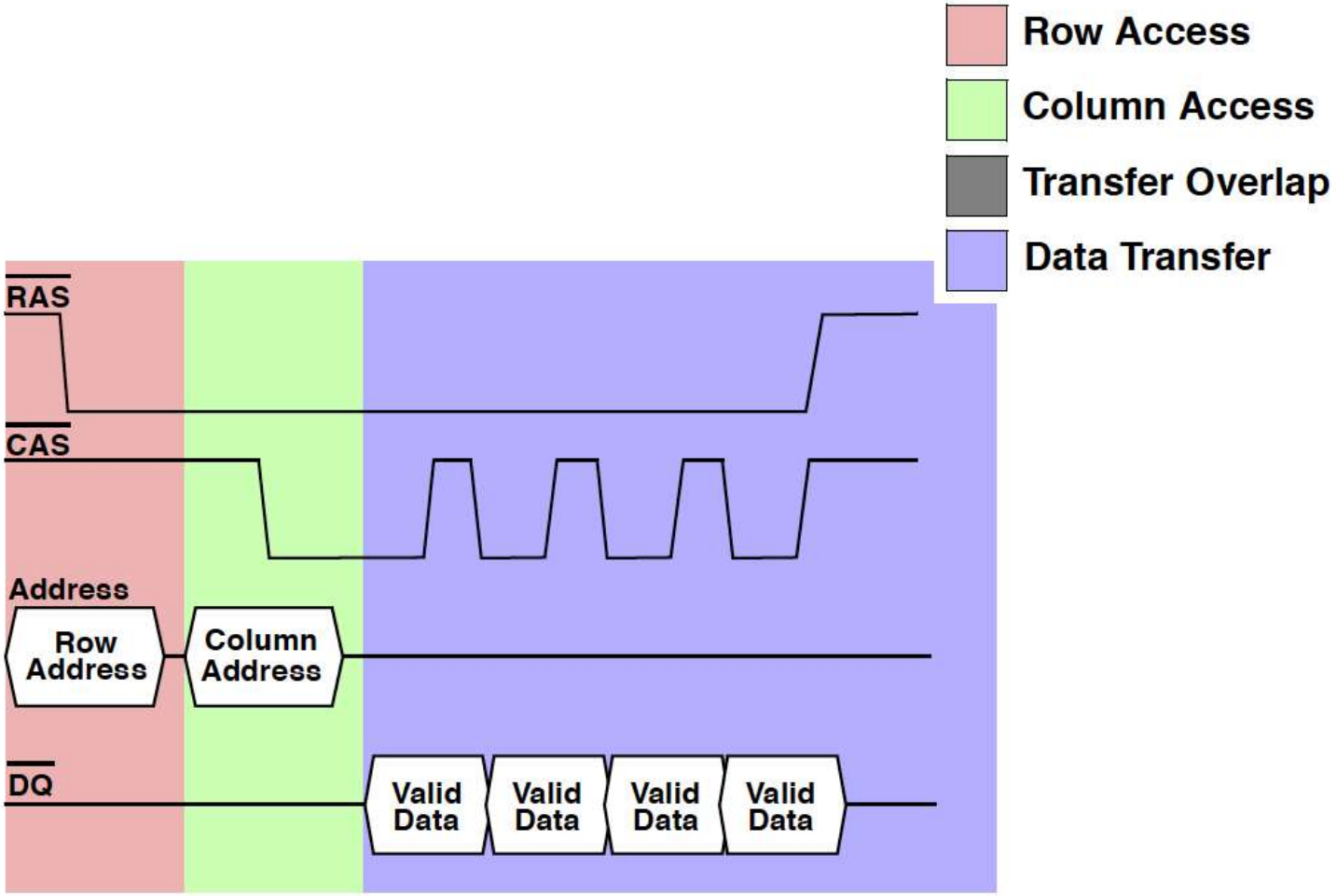
Read Timing for Extended Data Out (EDO)

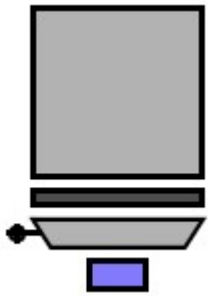




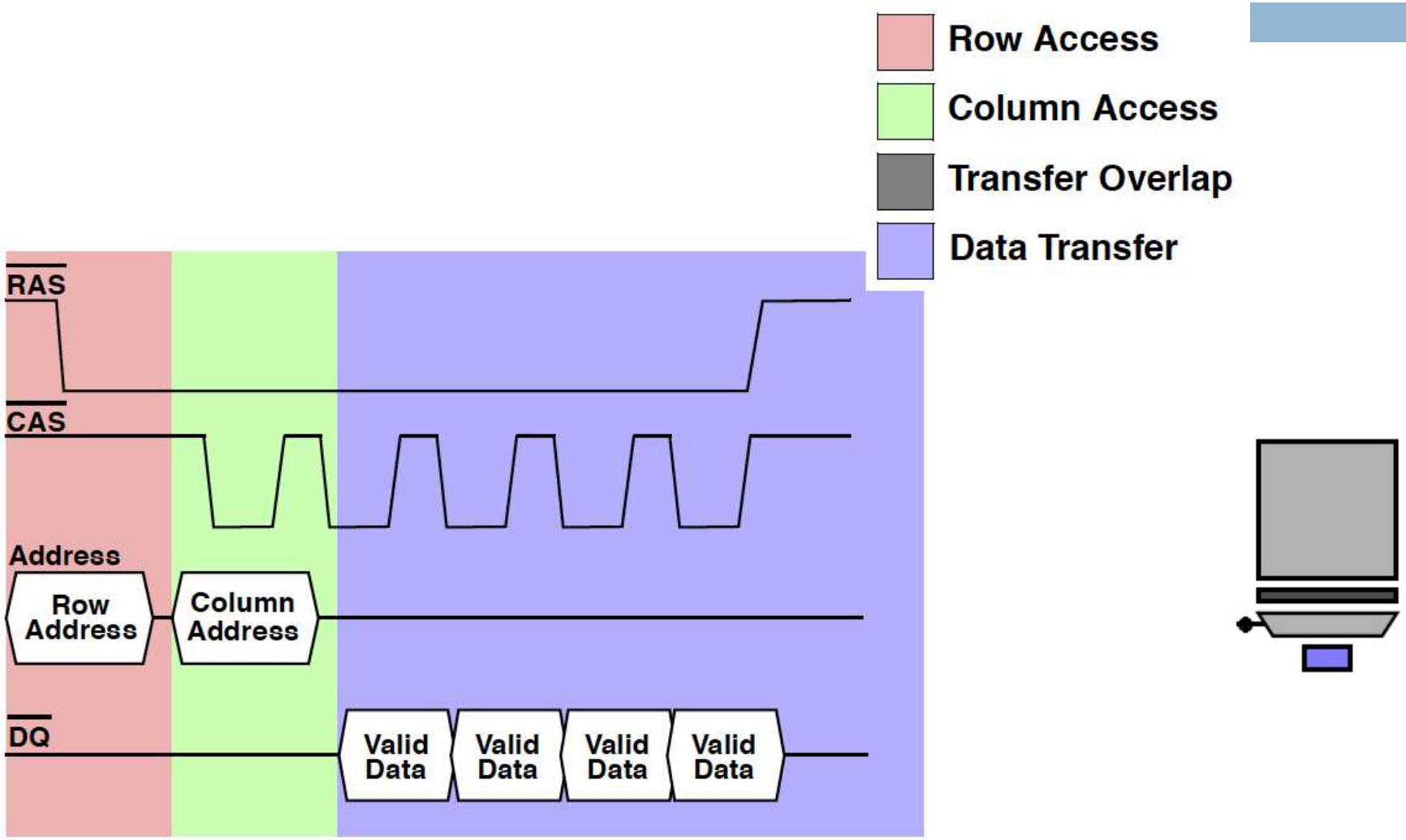
-- Assume that accesses are continuous, i.e., the next read address is a sequel of the current one.

Read Timing for Burst EDO

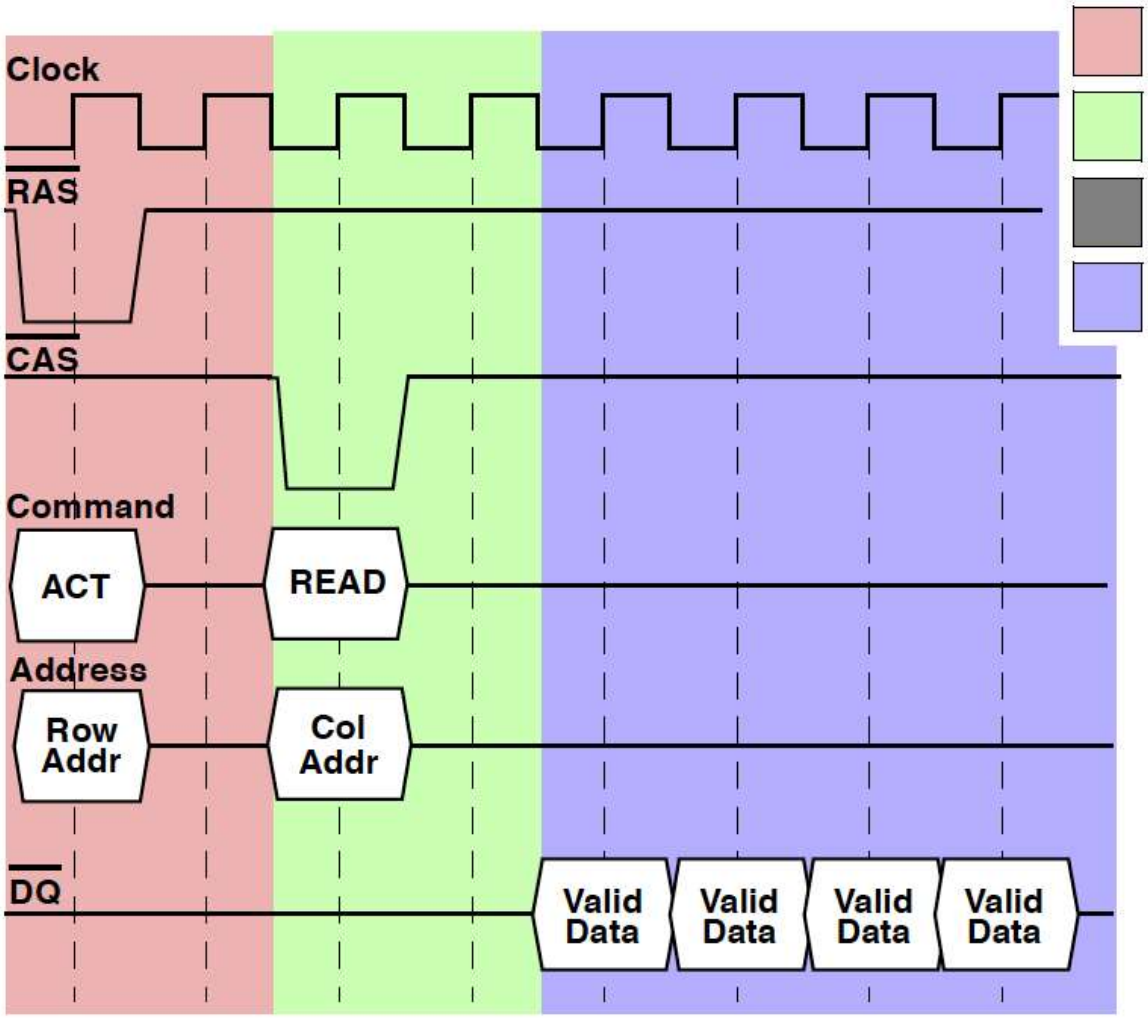


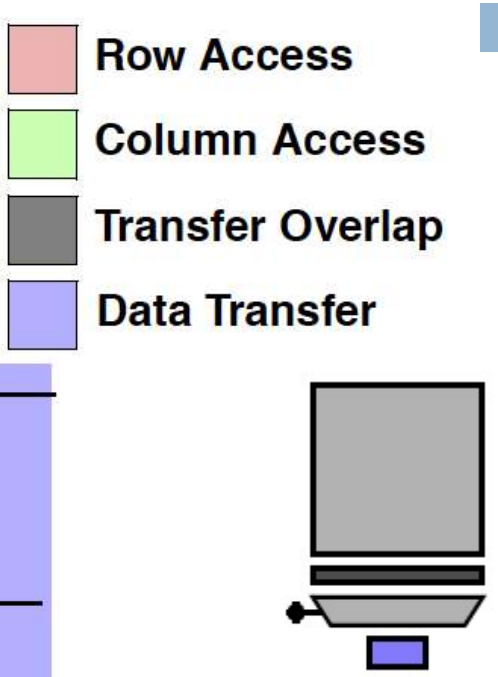


Read Timing for Pipeline Burst EDO



Read Timing for Synchronous DRAM





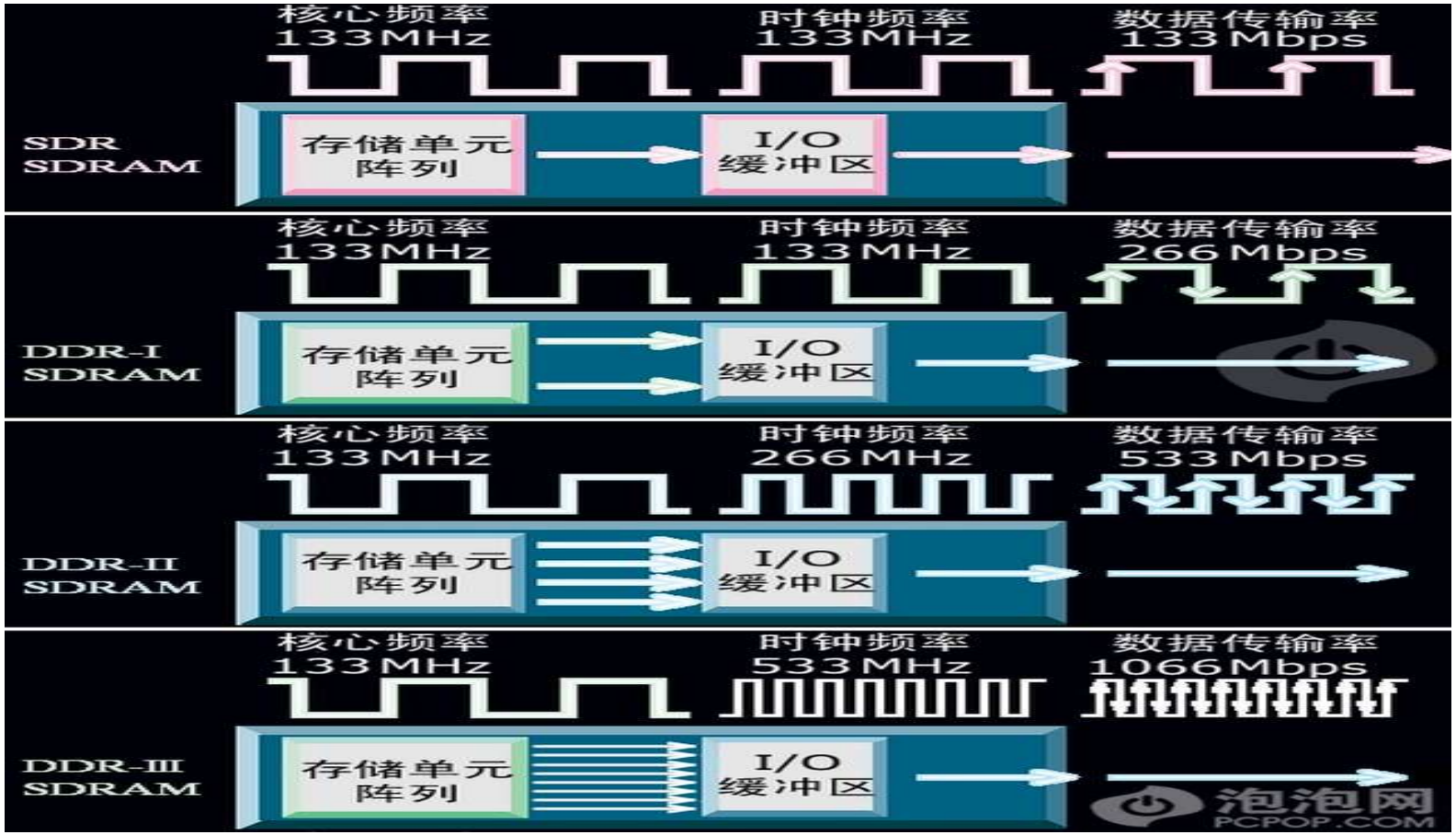
- Add synchronous dual buffers
- Interleave way in accessing banks
- Pipelined read or pipelined write

DDR SDRAM

- Double Data Rate SDRAM
- Data can be transmitted on both clock edges

DDR SDRAM Standard	Bus clock (MHz)	Internal rate (MHz)	Prefetch (min burst)	Transfer Rate (MT/s)	Voltage	DIMM pins	SO-DIMM pins	MicroDIMM pins
DDR	100–200	100–200	2n	200–400	2.5/2.6	184	200	172
DDR2	200–533	100–266	4n	400–1066	1.8	240	200	214
DDR3	400–1066	100–266	8n	800–2133	1.5	240	204	214
DDR4	800–1200	200–300	8n	1600–2400	1.2	288	260	214

Clocking and Prefetch Buffering for Different Generations



Source: “記憶體10年技術演進史，系統顆粒DDR與顯示顆粒GDDR差在哪？" by Tandee on internet

GDDR (Graphics Double Data Rate)

版本	GDDR	GDDR2	gDDR2	GDDR3	gDDR3	GDDR4	GDDR5
預取量	2bit	4bit	4bit	4bit	8bit	8bit	8bit
對應世代	DDR	DDR2	DDR2	DDR2	DDR3	DDR3	DDR3
Burst	2/4/8bit	4/8bit	4/8bit	4/8bit	4/8bit	4/8bit	8bit
額定電壓	2.5V	2.5V	1.8V	1.8V	1.5V	1.5V	1.5V
單顆容量	16/32MB	32MB	32/64/128MB	32/64/128MB	64/128MB	64/128MB	64/128MB
介面頻寬	16/32bit	32bit	16bit	32bit	16bit	32bit	16/32bit
封裝針腳	66/144	144	84	136/144	96	136	170
Bank數量	2 / 4	4 / 8	4 / 8	4 / 8	8	8 / 16	8 / 16
等效時脈	300-900	800-1000	700-1200	1000-2600	1000-2000	2000-3000	3600-6000

Source: “記憶體10年技術演進史，系統顆粒DDR與顯示顆粒GDDR差在哪？" by Tandee on internet