CMPE 200
Computer Architecture & Design

Lecture 4.
Memory Hierarchy (5)

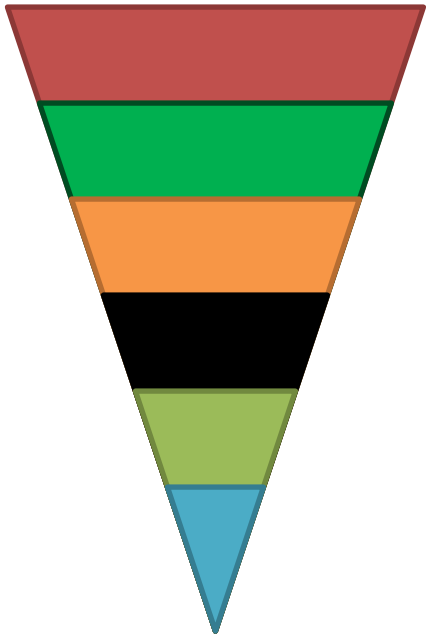
Haonan Wang

SAN JOSÉ STATE
UNIVERSITY

DRAM Subsystem Organization

Dram: Dynamic RAM

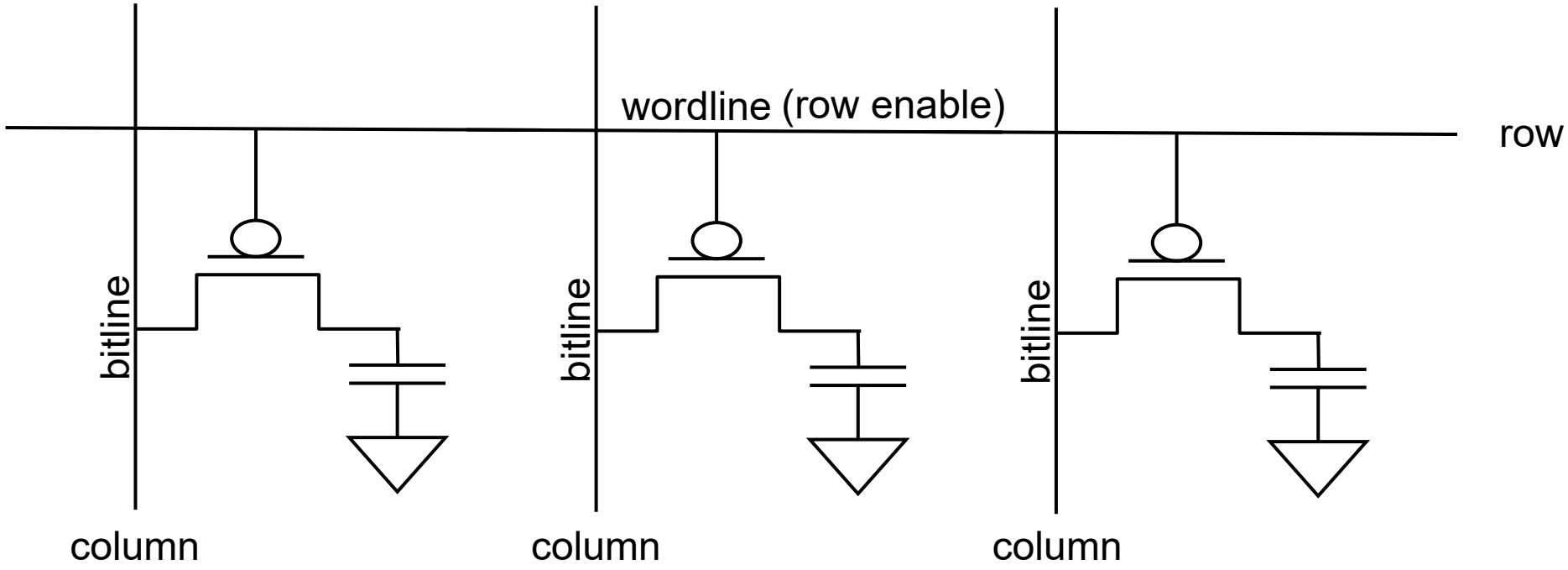
DRAM Organization:



Connected to an on-die memory controller

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

DRAM Cells



- **A DRAM cell consists of a capacitor and an access transistor**
- **It stores data in terms of charge in the capacitor**
- **Cheaper and larger than SRAM**
  - A DRAM chip consists of (10s of 1000s of) rows of such cells

DRAM Refresh

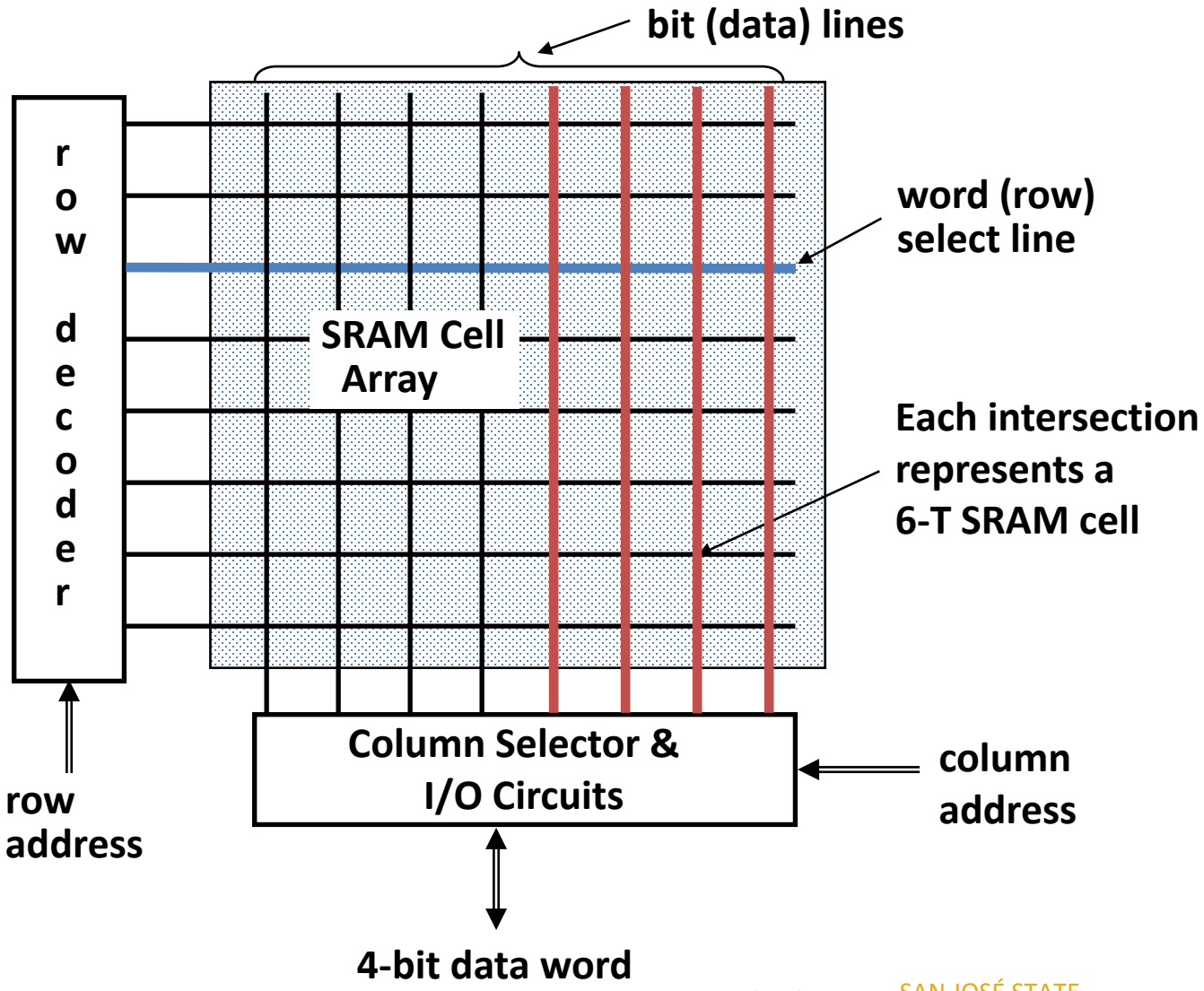
- **DRAM capacitor charge leaks over time**

- **The memory controller needs to refresh each row periodically to restore charge**
  - Dynamic (i.e., never in a stable state)
  - Activate each row every N ms (e.g., typical N = 64)

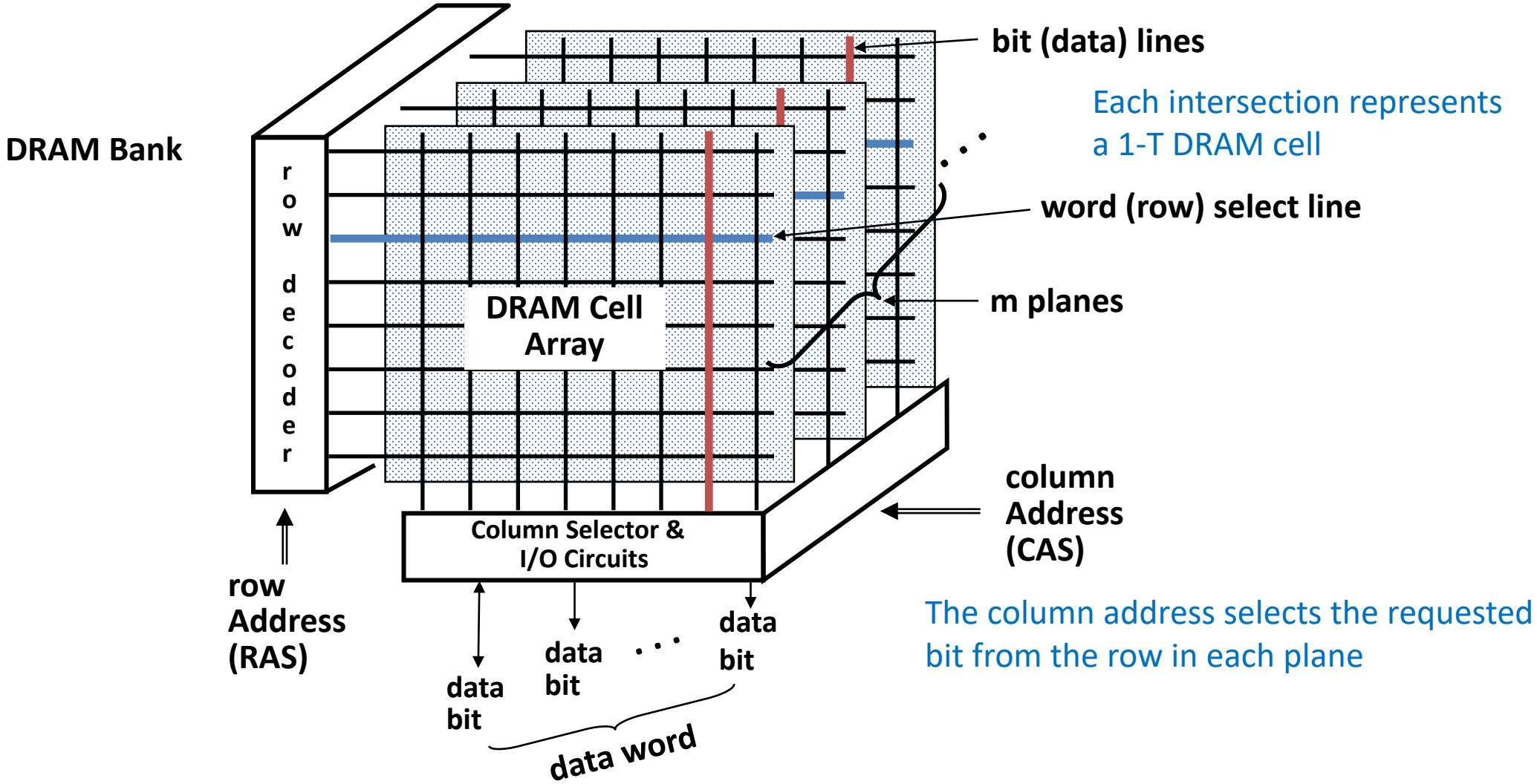
- **Downsides of refresh**
  - Energy consumption: Each refresh consumes energy
  - Performance degradation: DRAM rank/bank unavailable while refreshed
  - QoS/predictability impact: (Long) pause times during refresh
  - Refresh rate limits DRAM capacity scaling

Recall: SRAM Cache Design

- Each row holds a data block
- Column address selects the requested word from block



DRAM Design 1: The Classical



Each intersection represents a 1-T DRAM cell

The column address selects the requested bit from the row in each plane

DRAM Performance Metrics

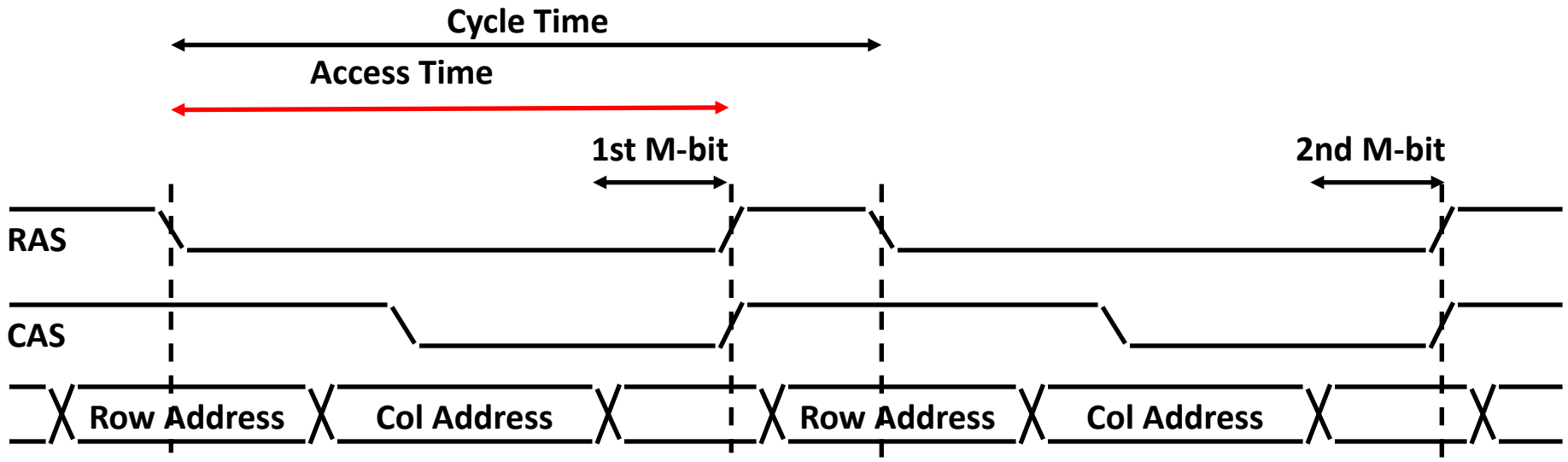
- **DRAM addresses are divided into 2 halves (row and column)**
  - *RAS* or *Row Access Strobe* that triggers the row decoder
  - *CAS* or *Column Access Strobe* that triggers the column selector

- **Latency: Time to access one word**
  - *Access Time*: time taken when word is read or written
    - read access and write access times can be different
  - *Cycle Time*: time between successive (read or write) requests

- **Bandwidth: How much data can be supplied per unit time**
  - Width of the data channel * channel usage frequency

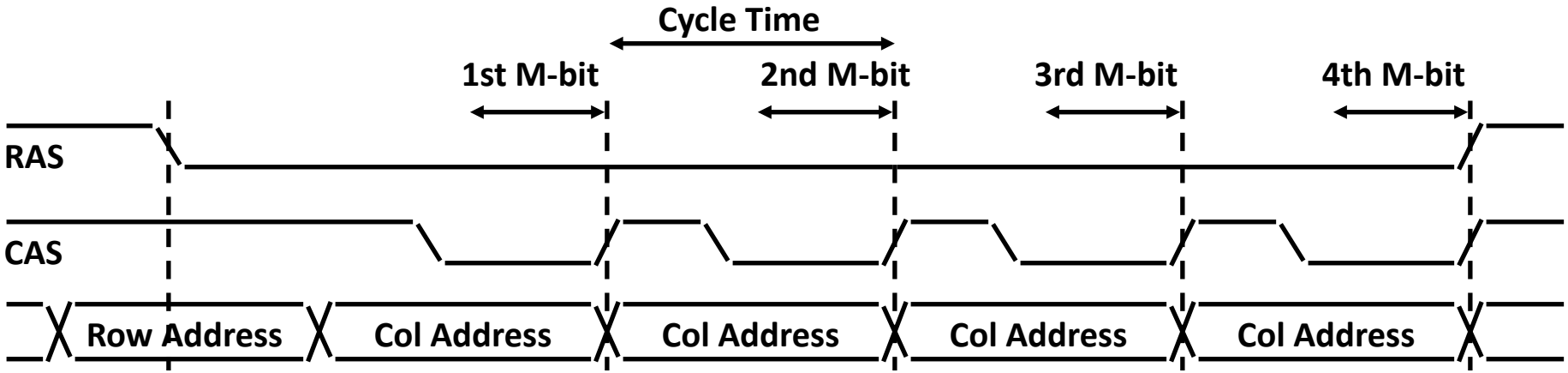
DRAM Design 1: The Classical

- **DRAM Organization:**
  - N rows x N column x M-bit (planes)
  - Reads or Writes M-bit at a time
  - Each M-bit access requires a RAS / CAS cycle



DRAM Design 1: The Classical

- **Page Mode: A row is kept "open" by keeping the RAS asserted**
  - Pulse CAS to access other M-bit blocks on **that** row
  - Successive reads or writes within the row are faster since don't have to precharge and (re)access that row



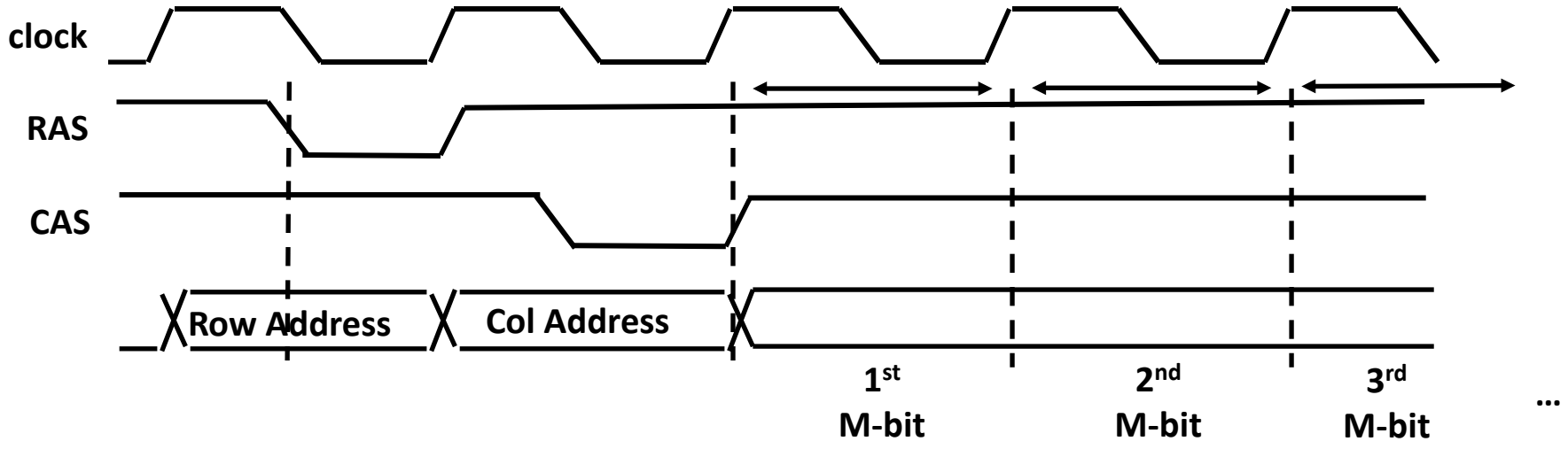
DRAM Design 2: Synchronous DRAMs

- **Like page mode DRAMs, synchronous DRAMs (SDRAMs) can transfer a burst of data from a series of sequential addresses in the same row**

- **For words in the same burst, don't have to provide the complete (row and column) addresses**
  - The entire row is loaded into a row buffer (SRAM).
  - Specify the starting (row+column) address and the burst length (burst must be in the same row).
  - Data words in the burst are then accessed from that SRAM under control of a clock signal.

Synchronous DRAM (SDRAM) Operation

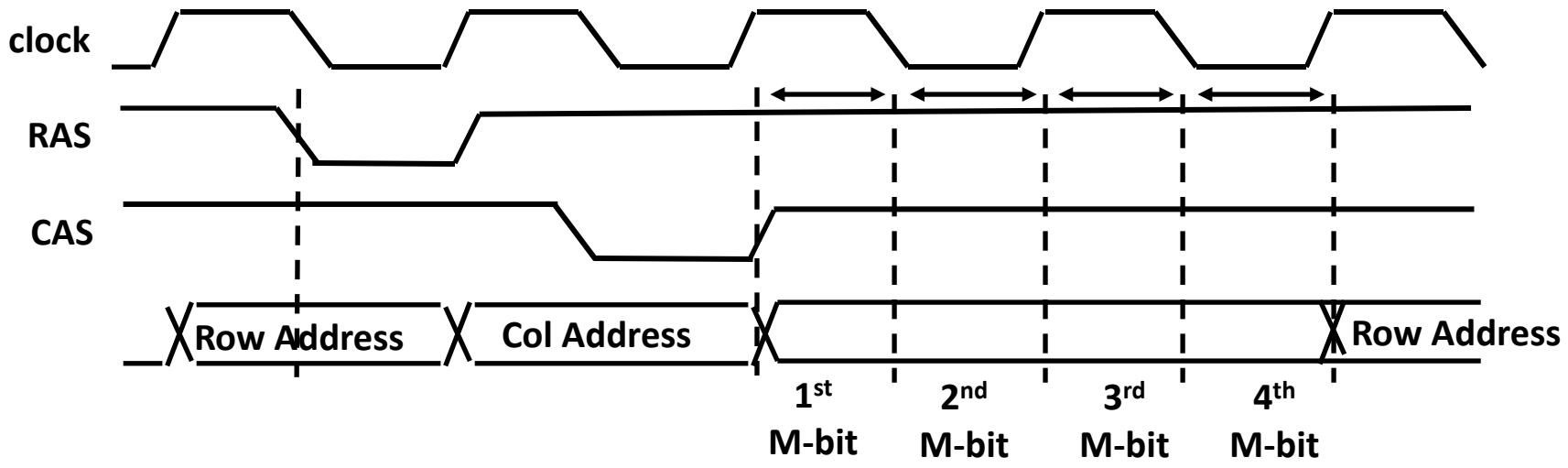
- **After RAS loads a row into the SRAM cache**
  - Input CAS as the starting "burst" address along with a burst length to read a burst of data from a series of sequential addresses within that row on the clock edge



http://en.wikipedia.org/wiki/DDR_SDRAM

DDR (Double Data Rate) SDRAMs

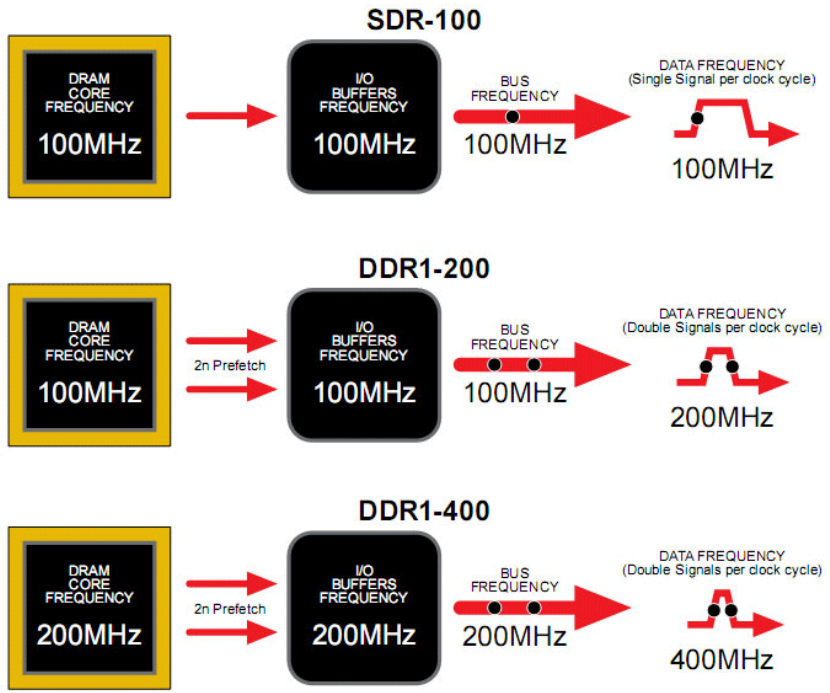
- **Transfers burst data on both the rising and falling edge of the clock (so twice fast)**
  - 2n core prefectch



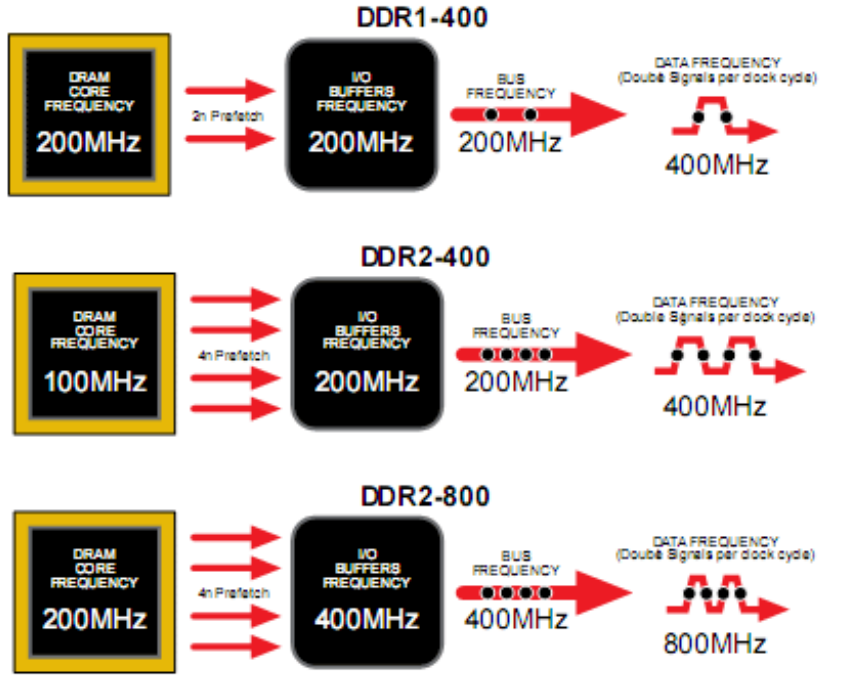
http://en.wikipedia.org/wiki/DDR_SDRAM

DDR (Double Data Rate) SDRAMs

- **DDR1 VS DDR1+:**



▲ Simplified Comparison between SDR-100, DDR1-200 and DDR1-400
Illustration: Ryan J. Leng



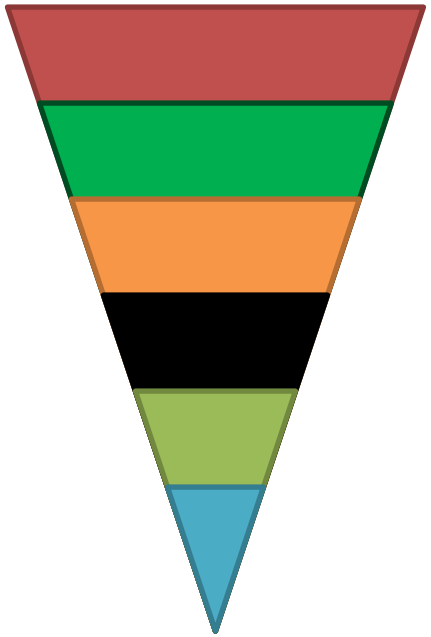
▲ Simplified Comparison between DDR1-400, DDR2-400 and DDR2-800
Illustration: Ryan J. Leng

- **DDR2 vs. QDR**

DRAM Subsystem Organization

Dram: Dynamic RAM

DRAM Organization:



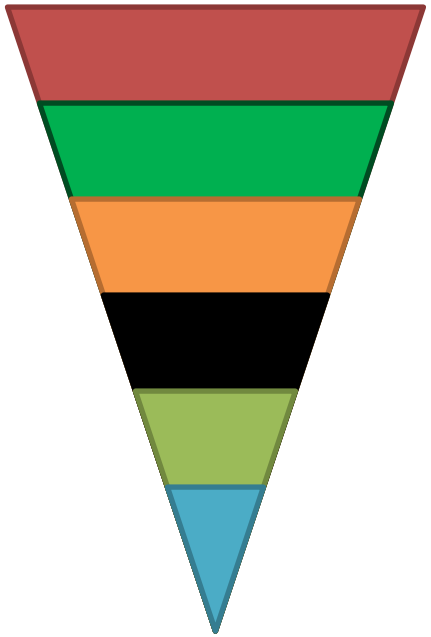
- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/Column

- **Chip:**
  - Consists of multiple banks (2-16 in Synchronous DRAM)
    - Banks works in parallel to overlap delay
  - Banks share command/address/data buses
  - The chip itself has a narrow interface (4-16 bits per read)

DRAM Subsystem Organization

Dram: Dynamic RAM

DRAM Organization:



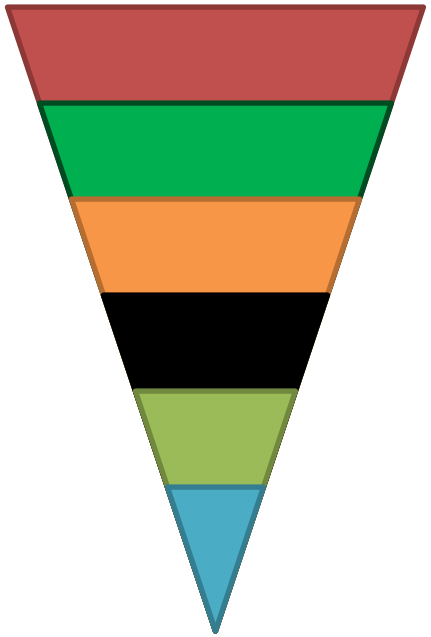
- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/Column

- **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
  - E.g., Using 8 chips with 8-bit interface to form a 64-bit wide Rank

DRAM Subsystem Organization

Dram: Dynamic RAM

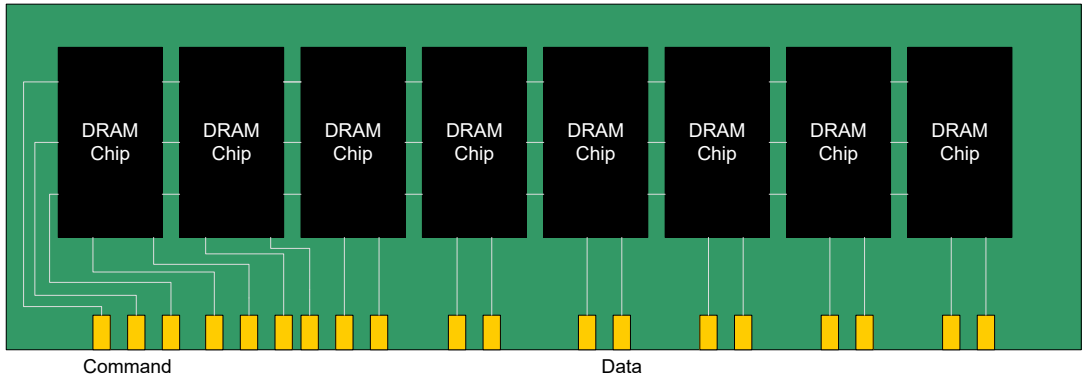
DRAM Organization:



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

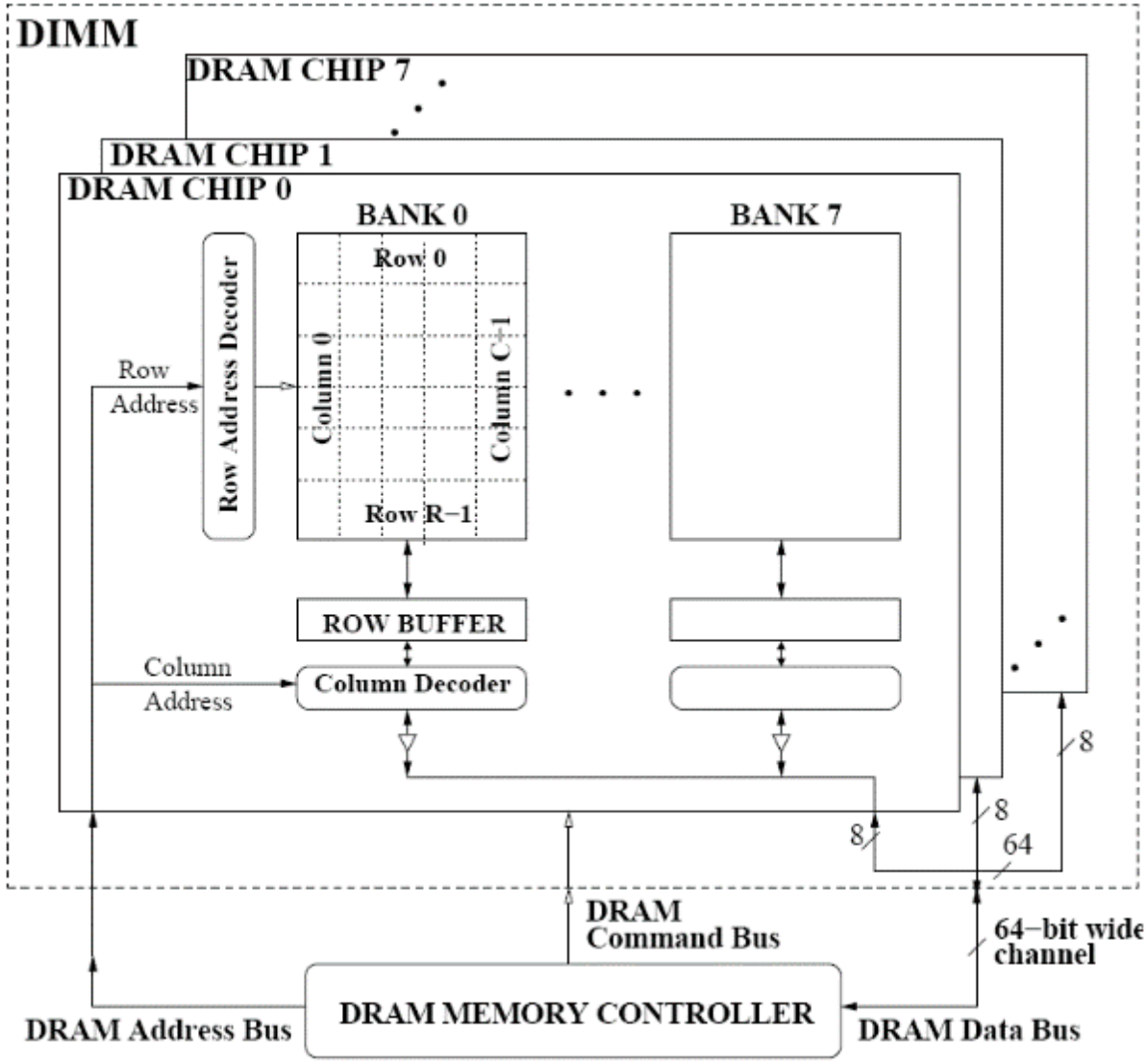
- **DIMM (dual inline memory module):**
  - A DRAM module consists of one or more ranks
  - This is what you plug into your motherboard

DIMM (Dual Inline Memory Module)



- **Contains DRAM chips each have data widths of x4 or x8**
  - Can be on both sides
- **Can have more than one rank**
  - Increase storage capacity
  - Only one rank accessible at a time
- **SIMM vs DIMM**
  - DIMM has separate contacts on each side of the board to provides twice as much data rate.

A 64-bit Wide DIMM (One Rank)

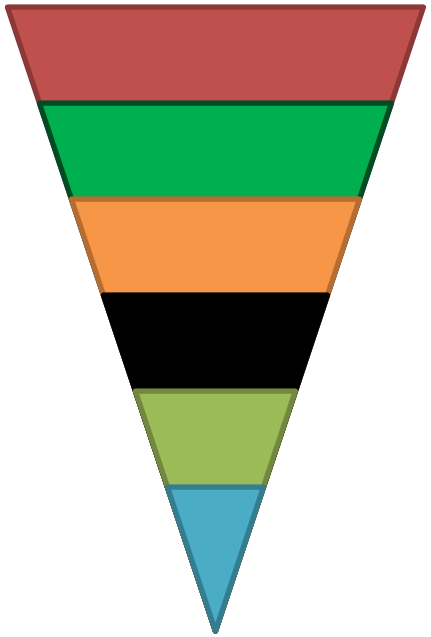


- **Advantages:**
  - Acts like a high-capacity DRAM chip with a wide interface
  - Flexibility: memory controller does not need to deal with individual chips
- **Disadvantages:**
  - Granularity: Accesses cannot be smaller than the interface width

DRAM Subsystem Organization

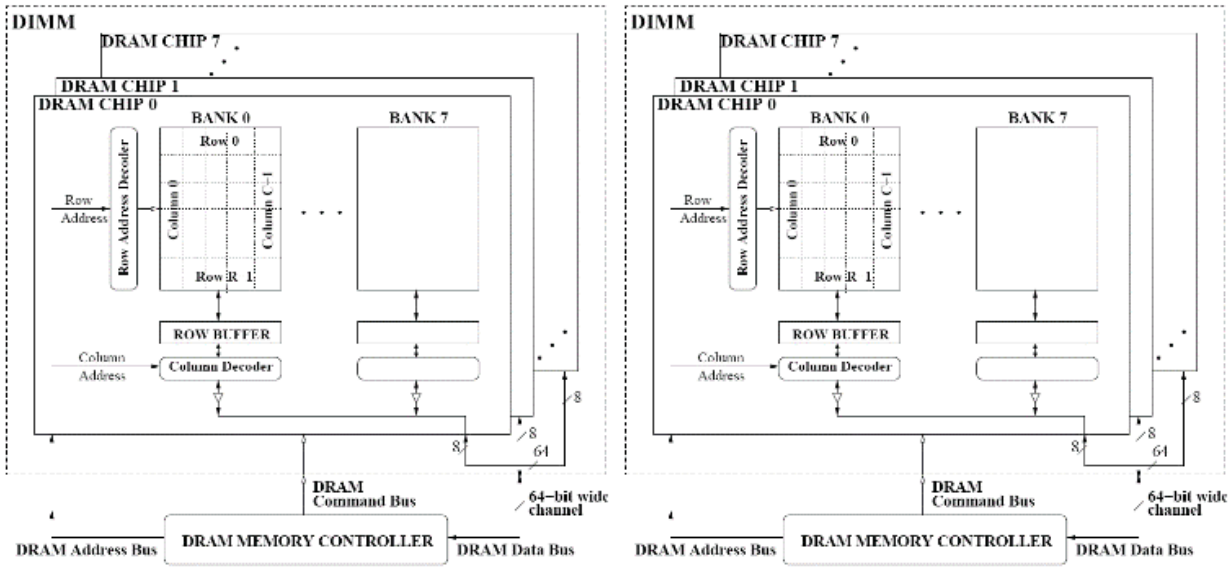
Dram: Dynamic RAM

DRAM Organization:

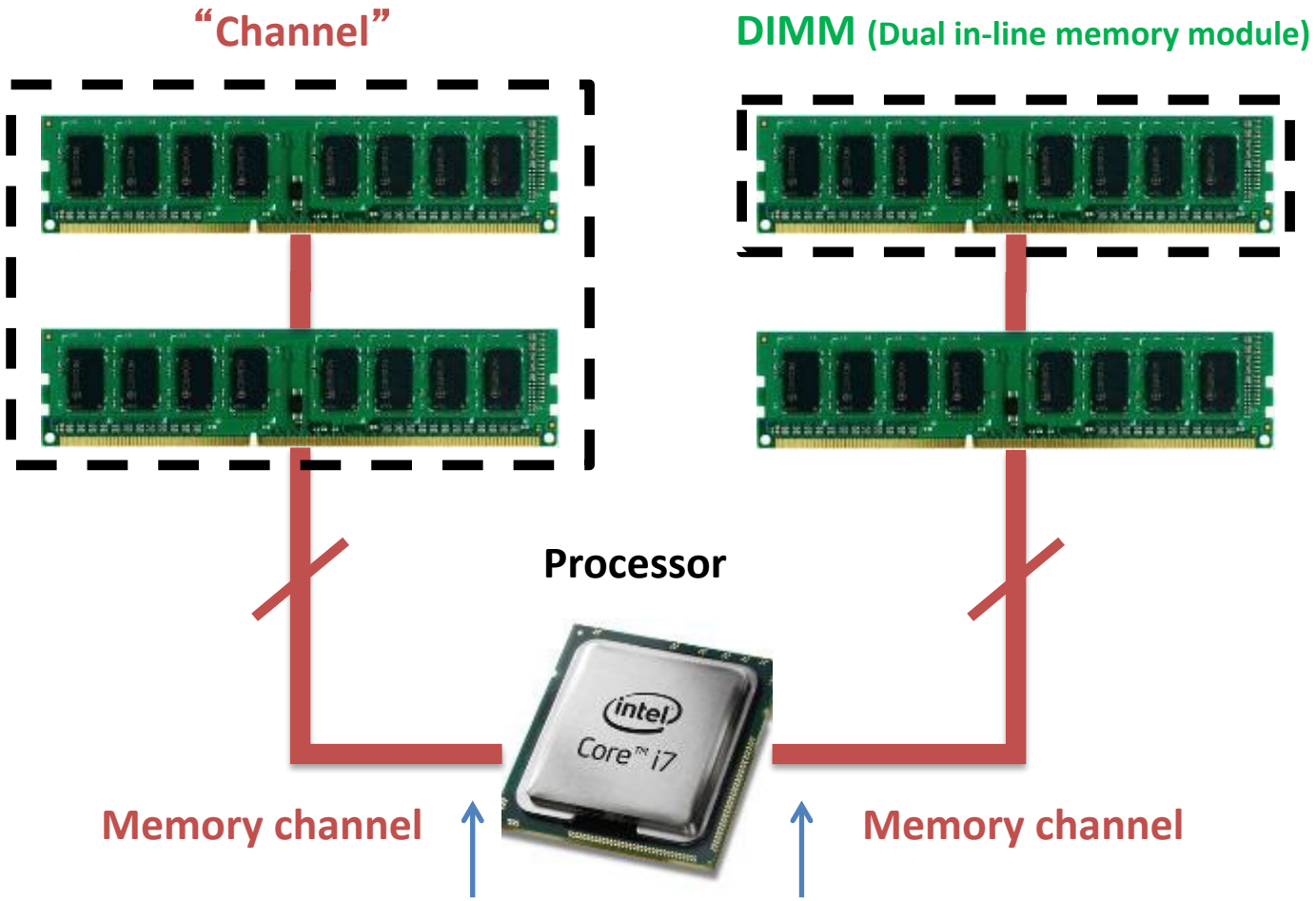


- Channel
- DIMM
- Rank
- Chip
- Bank
- Row/Column

- **Channel: Independent memory subsystem**
  - E.g., 2 independent Channels:

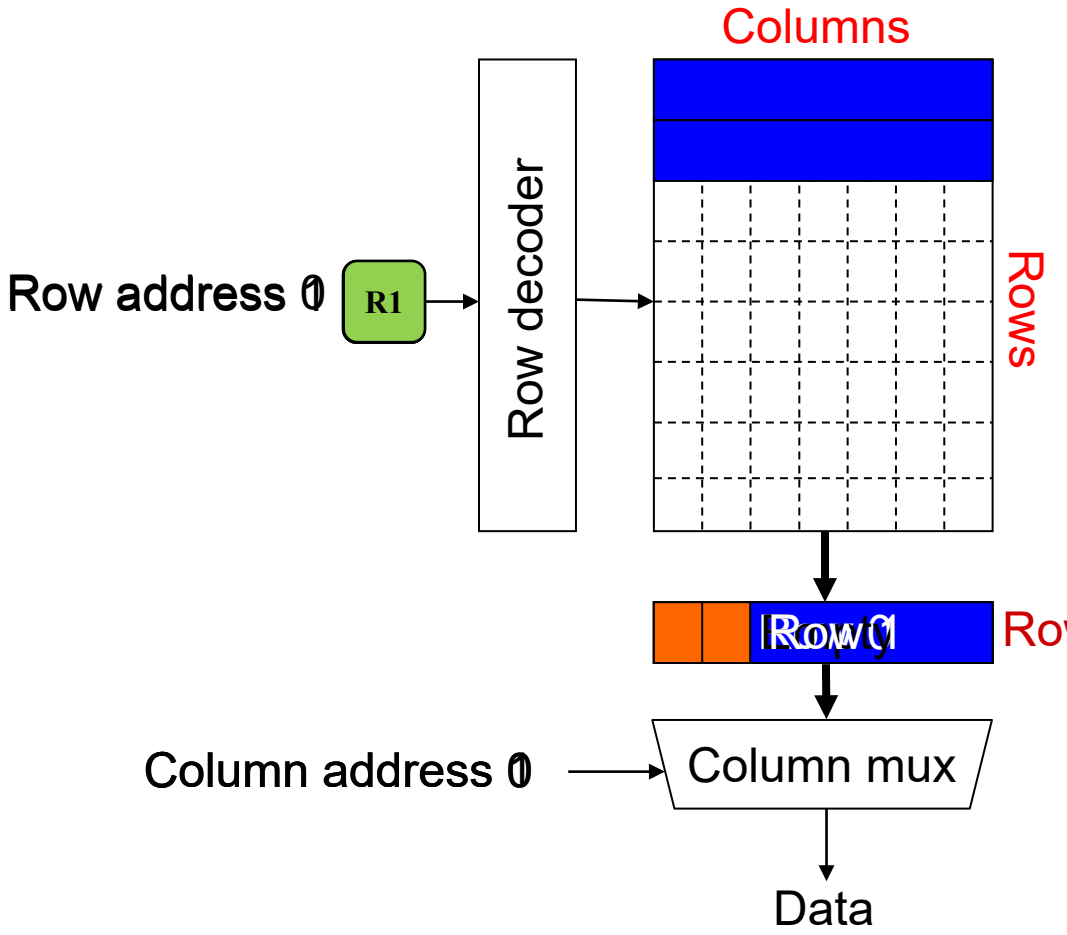


The DRAM subsystem



Each channel is connected to its own on-die memory controller

Row Operations & Row Buffer Locality



Access Address:	Row Operation:
RBL=2 (Row 0, Column 0)	Activation
RBL=2 (Row 0, Column 1)	No operation
RBL=1 (Row 1, Column 0)	Restore, Precharge, Activation

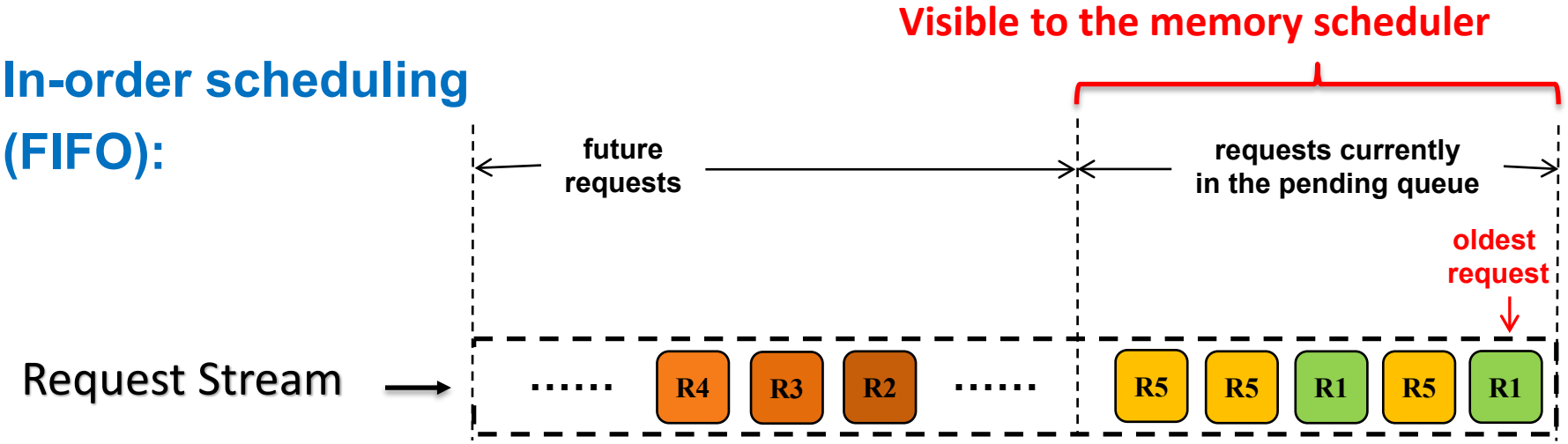
CONFLICT !

Improving Row Buffer Locality (RBL) is the key to improve DRAM efficiency

Credit: Onur Mutlu.

RBL & Memory Scheduling Schemes

In-order scheduling (FIFO):



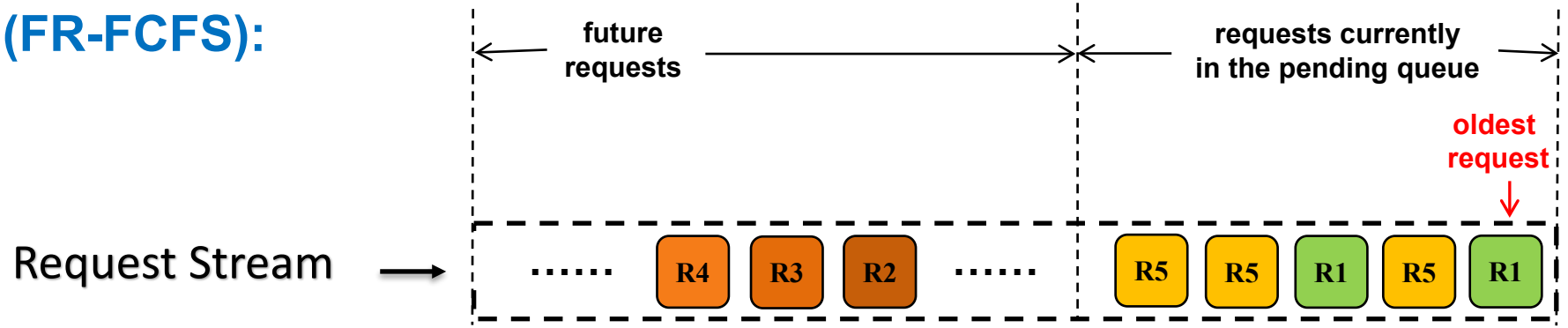
Activation Counter:

R1: Activation = 1
R5: Activation = 2
R1: Activation = 3
R5: Activation = 4
R5: Activation = 4

Same activation

Avg RBL = 5 / 4 = 1.25

Out-of-order scheduling (FR-FCFS):



Activation Counter:

R1: Activation = 1
R1: Activation = 1

Same activation

R5: Activation = 2
R5: Activation = 2
R5: Activation = 2

Same activation

Avg RBL = 5 / 2 = 2.5

DRAM Milestones

	DRAM	Page DRAM	Page DRAM	Page DRAM	SDRAM	DDR SDRAM
Module Width	16b	16b	32b	64b	64b	64b
Year	1980	1983	1986	1993	1997	2000
Mb/chip	0.06	0.25	1	16	64	256
Die size (mm^2)	35	45	70	130	170	204
Pins/chip	16	16	18	20	54	66
BWidth (MB/s)	13	40	160	267	640	1600
Latency (nsec)	225	170	125	75	62	52

- **In the time that the memory to processor bandwidth has more than doubled the memory latency has improved by a factor of only 1.2 to 1.4**

Patterson, CACM Vol 47, #10, 2004

Review: DRAM vs. SRAM

- **DRAM**
  - Slower access (capacitor)
  - Higher density (1T 1C cell)
  - Lower cost
  - Requires refresh (power, performance, circuitry)
  - Manufacturing requires putting capacitor and logic together

- **SRAM**
  - Faster access (no capacitor)
  - Lower density (6T cell)
  - Higher cost
  - No need for refresh
  - Manufacturing compatible with logic process (no capacitor)

Conclusion Time

What are the levels in the DRAM organization?

Channel, DIMM, Rank, Chip, Bank, Row, Column

What is the idea behind having multiple memory banks?

Hiding latency

What is row buffer?

Cache for DRAM rows

SAN JOSÉ STATE UNIVERSITY *powering* SILICON VALLEY