# CS230: Digital Logic Design and Computer Architecture

Lecture 21: DRAM Organization and Controller

https://www.cse.iitb.ac.in/~biswa/courses/CS230/main.html

https://www.cse.iitb.ac.in/~biswa/

DRAM Organization

Channel

DIMM

Rank

Chip

Bank

Row

Column

Rank 1 with 8 chips

Rank 0 with 8 chips

Computer Architecture

2

Rank

DIMM (Dual in-line memory module)

SIDE

4.00

Front of DIMM

Back of DIMM

SPD
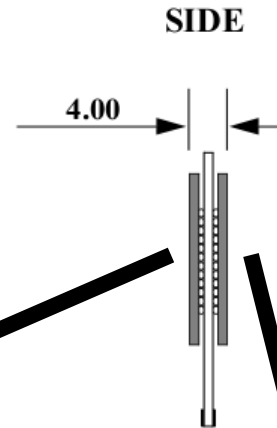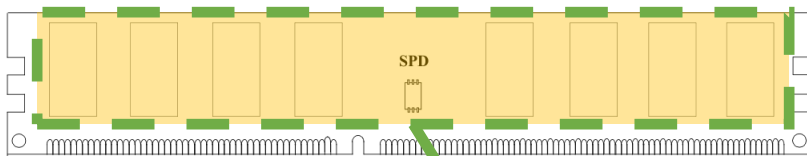
Rank 0: collection of 8 chips

Rank 1

Computer Architecture

Ranks, Banks, Rows, Columns

Computer Architecture

Ranks, Banks, Rows, Columns

Rank 0

64 bits

16
16
16
16

Chip

Bank

Row

Column

Bank 0   Bank 1

Bank 2   Bank 3
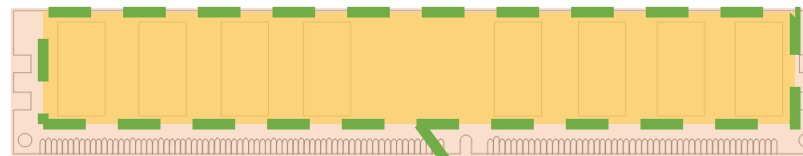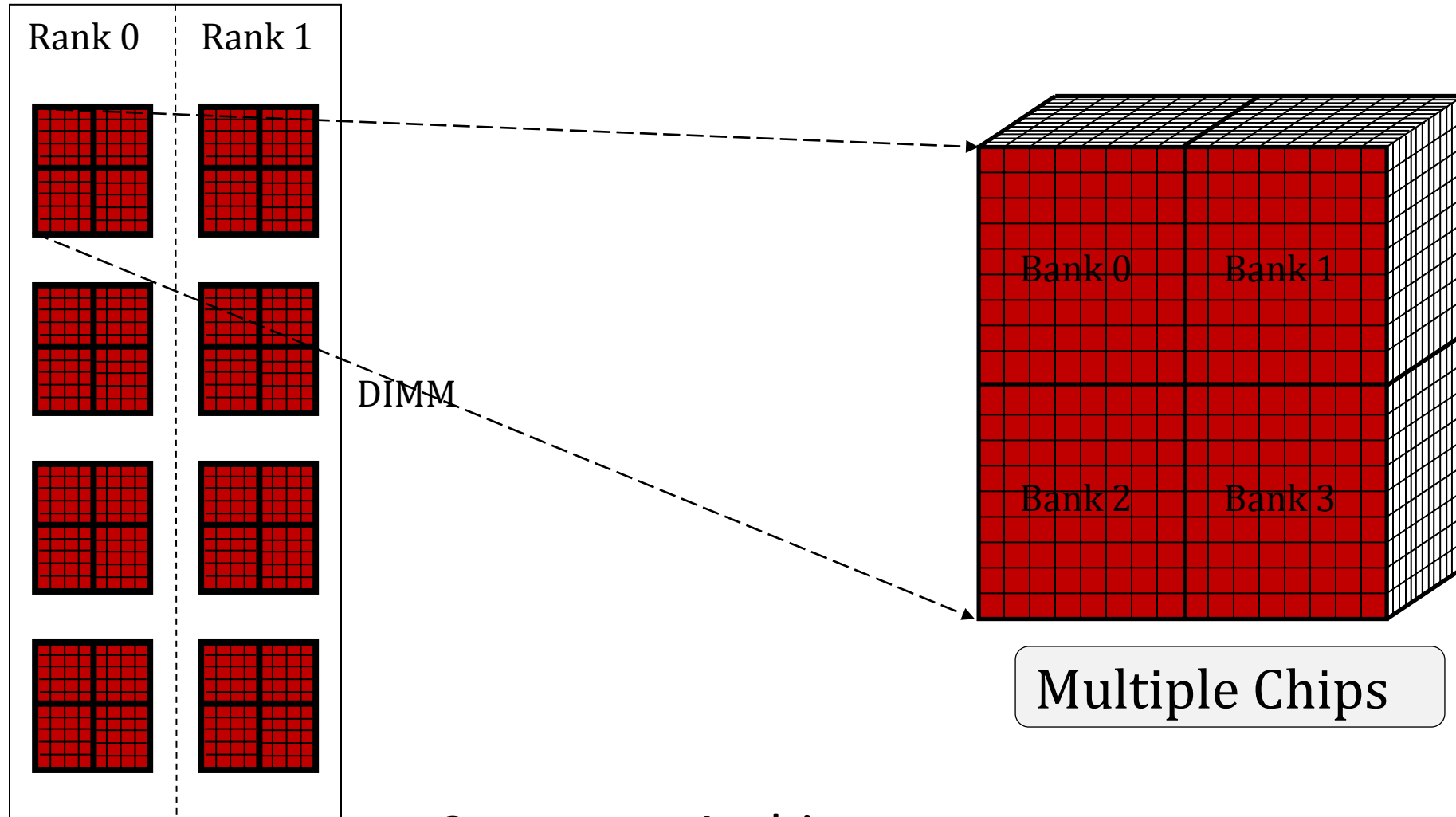
X16 device (DRAM width, # output pins)
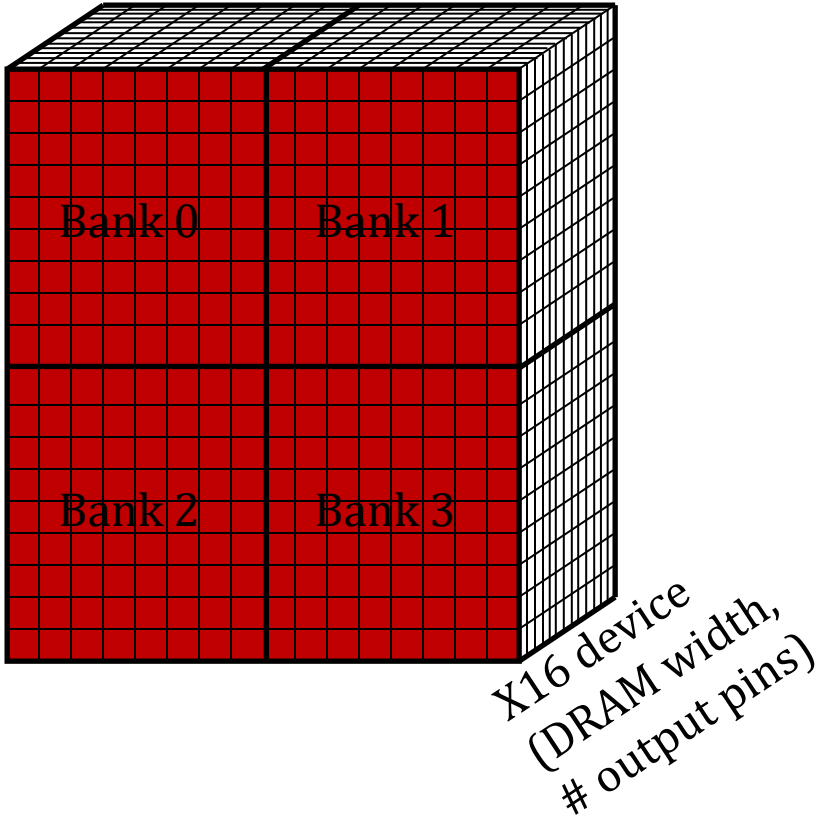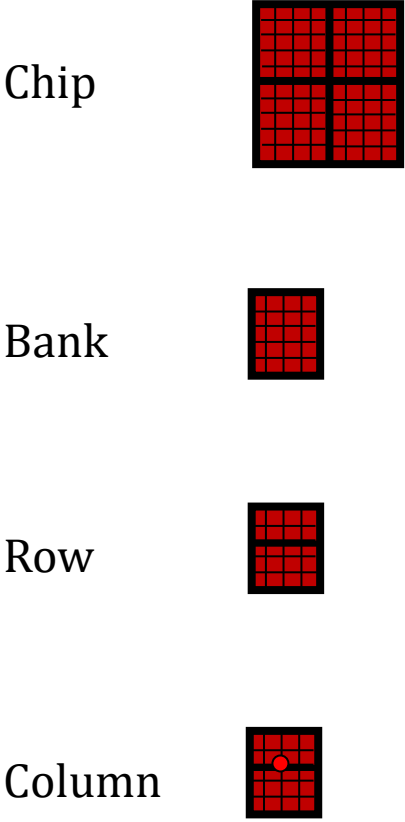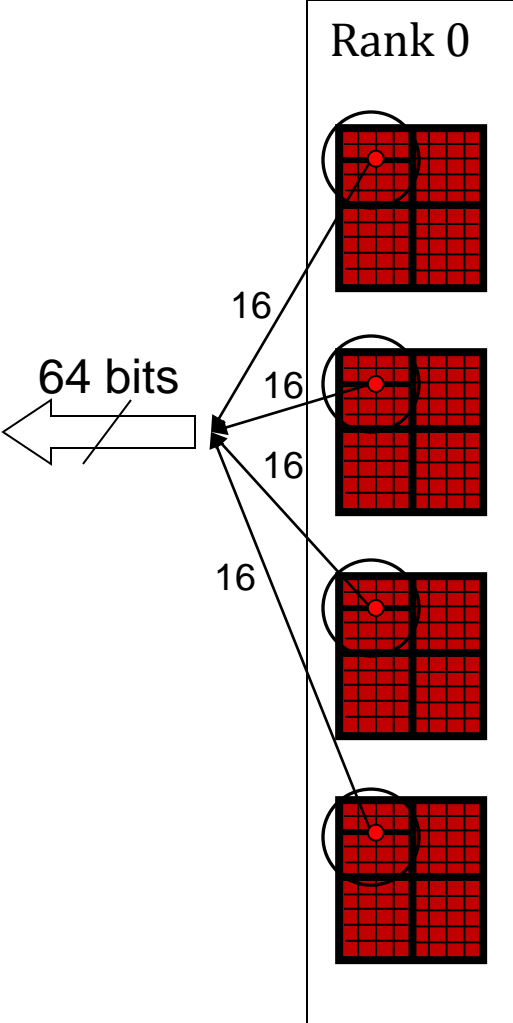
16-bit interface: 16 bits from each chip in one go

Ranks, Banks, Rows, Columns

Each rank has 64-bit wide data bus

If a rank is of width x8 then # DRAM chips ??

What about x4, # DRAM chips ??

If a rank is of width x8 then # DRAM chips ?? 8

What about x4, # DRAM chips ?? 16

Computer Architecture

Row Decoder and Row buffer (Sense Amplifier)

2KB

Row Decoder

Row Buffer

Column mux
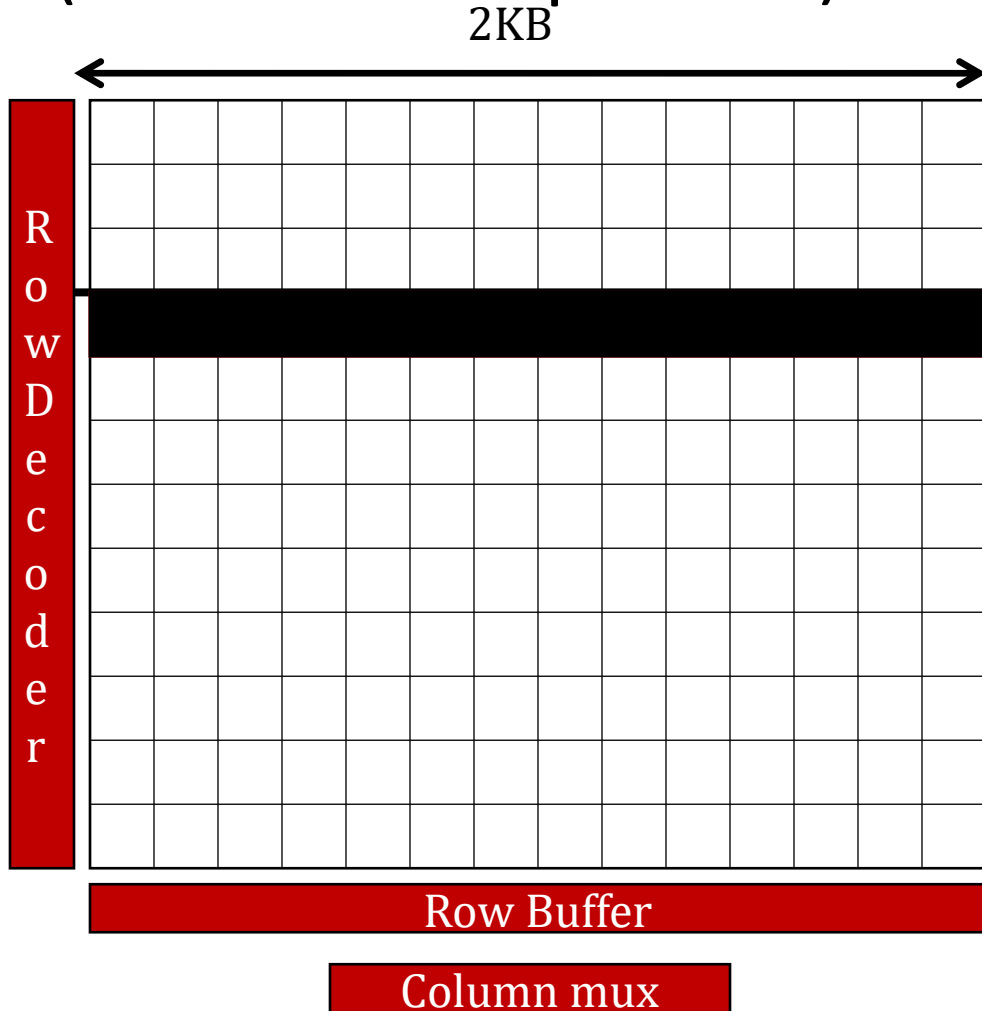
Each bank has a row buffer

Stores the last used row

Computer Architecture

An Example

2Gb * 8 DRAM Chips (one side of the rank)

Total 16 chips + 2 chips for ECC (for both the ranks)

64 bit + 8 bit ECC interface (72 bit wide DIMM)

Transferring a 64B cache line will take 8 transfers of 8B each

8B will come from 8 chips (8 bits from one chip)

1 bit from each DRAM array assuming 8 DRAM arrays per bank

# DRAM to LLC interaction

Physical memory space

# DRAM to LLC interaction

Physical memory space
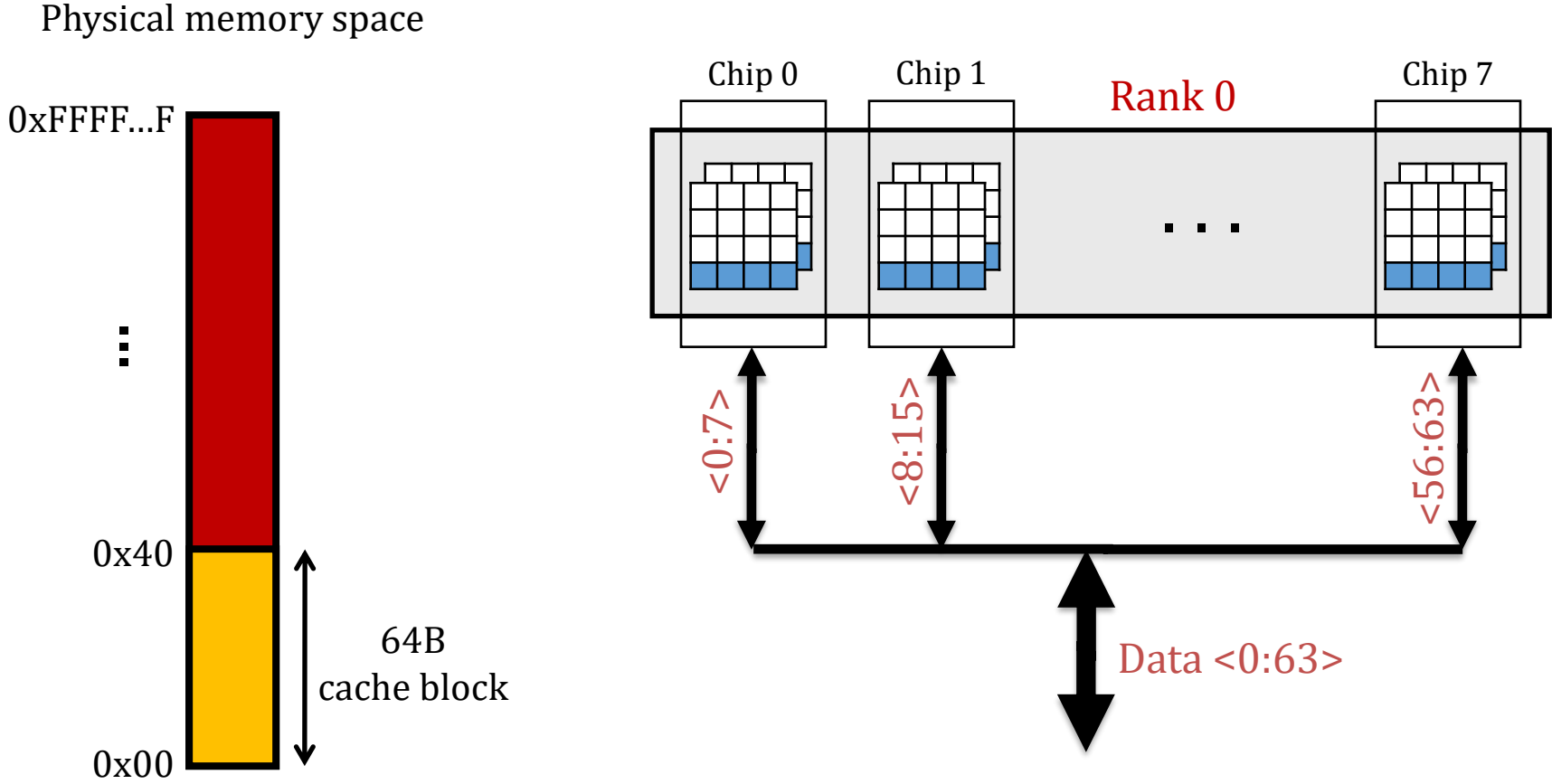
DRAM to LLC interaction
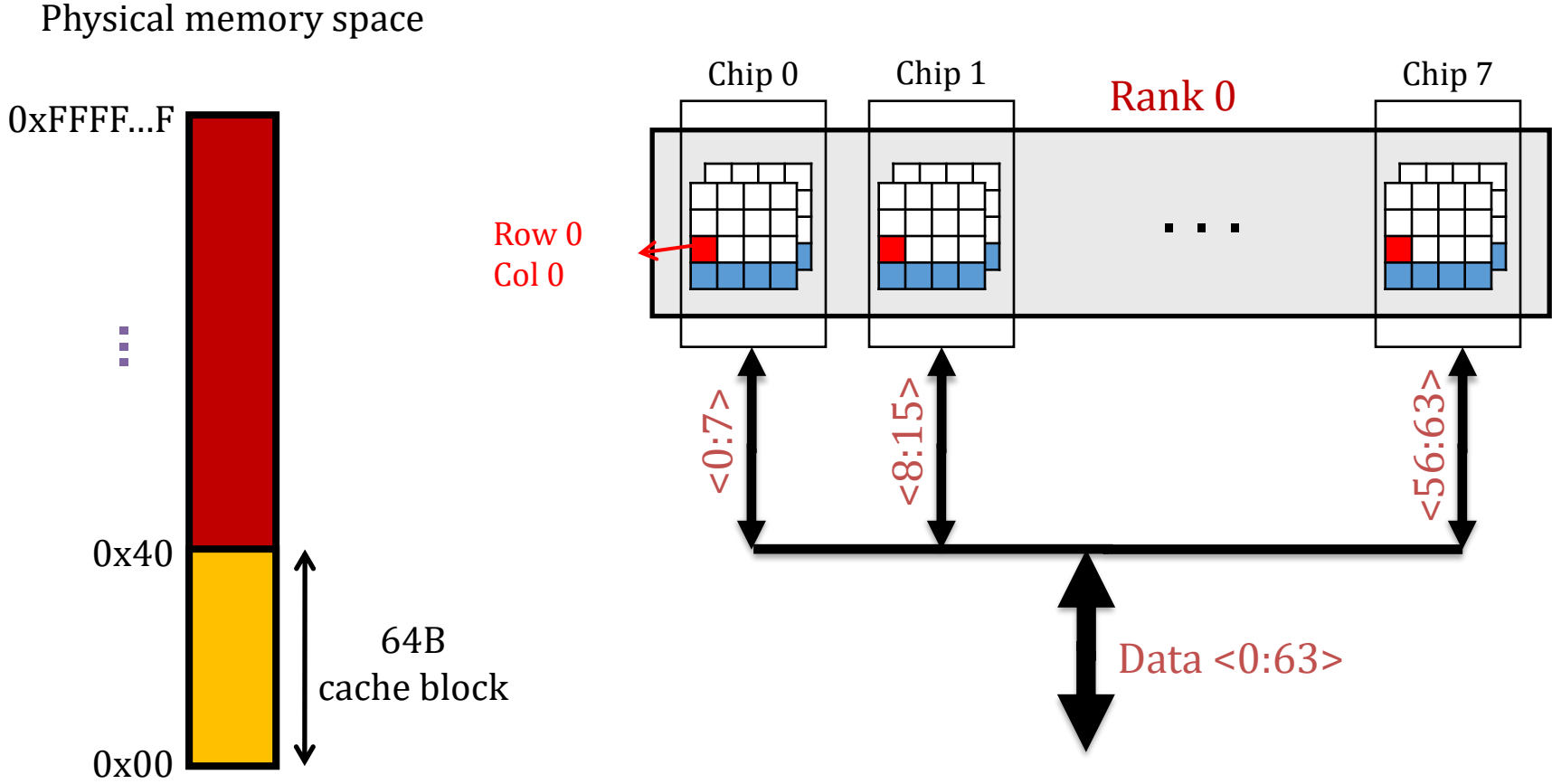
Physical memory space
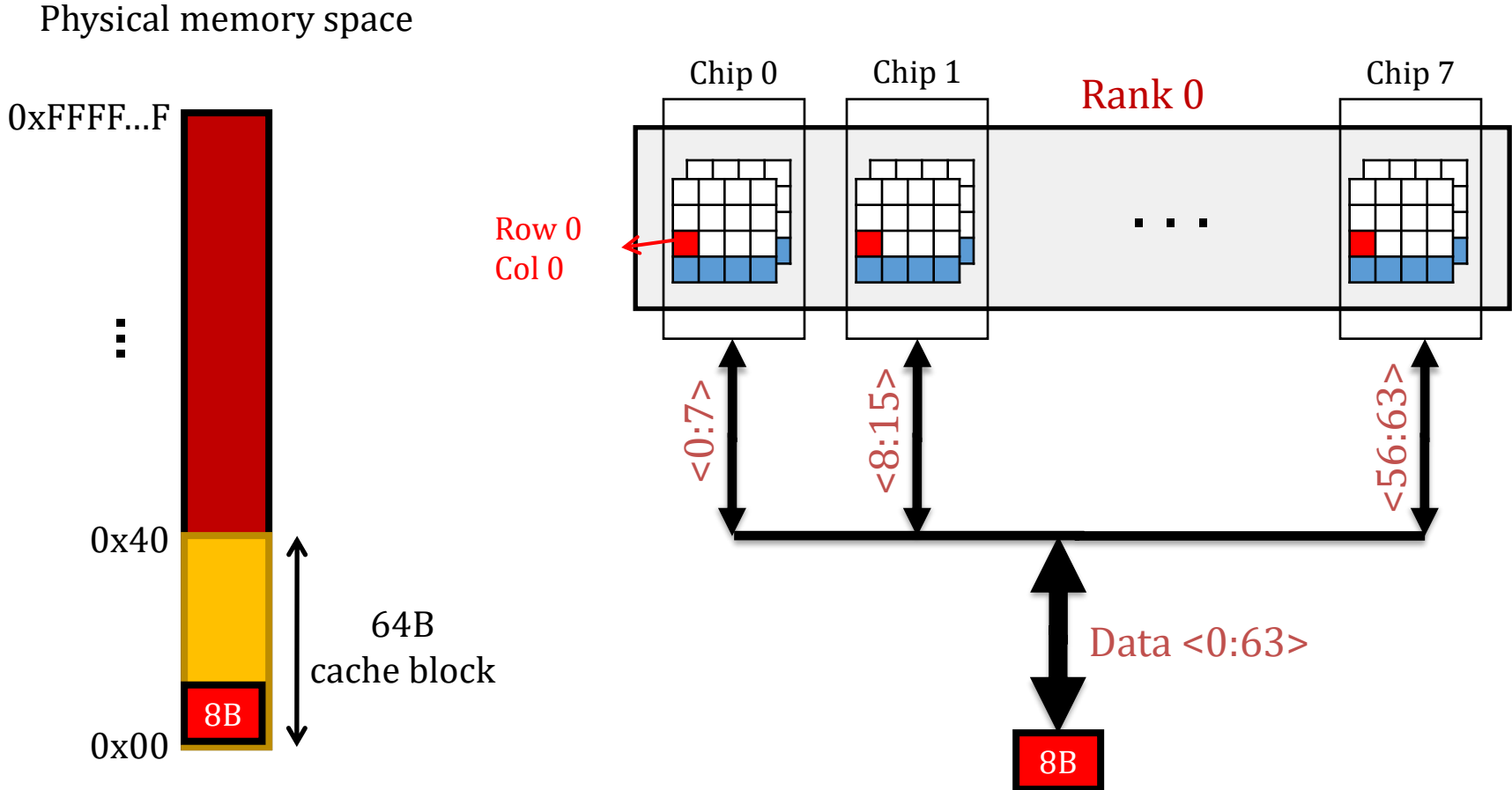
DRAM to LLC

Physical memory space
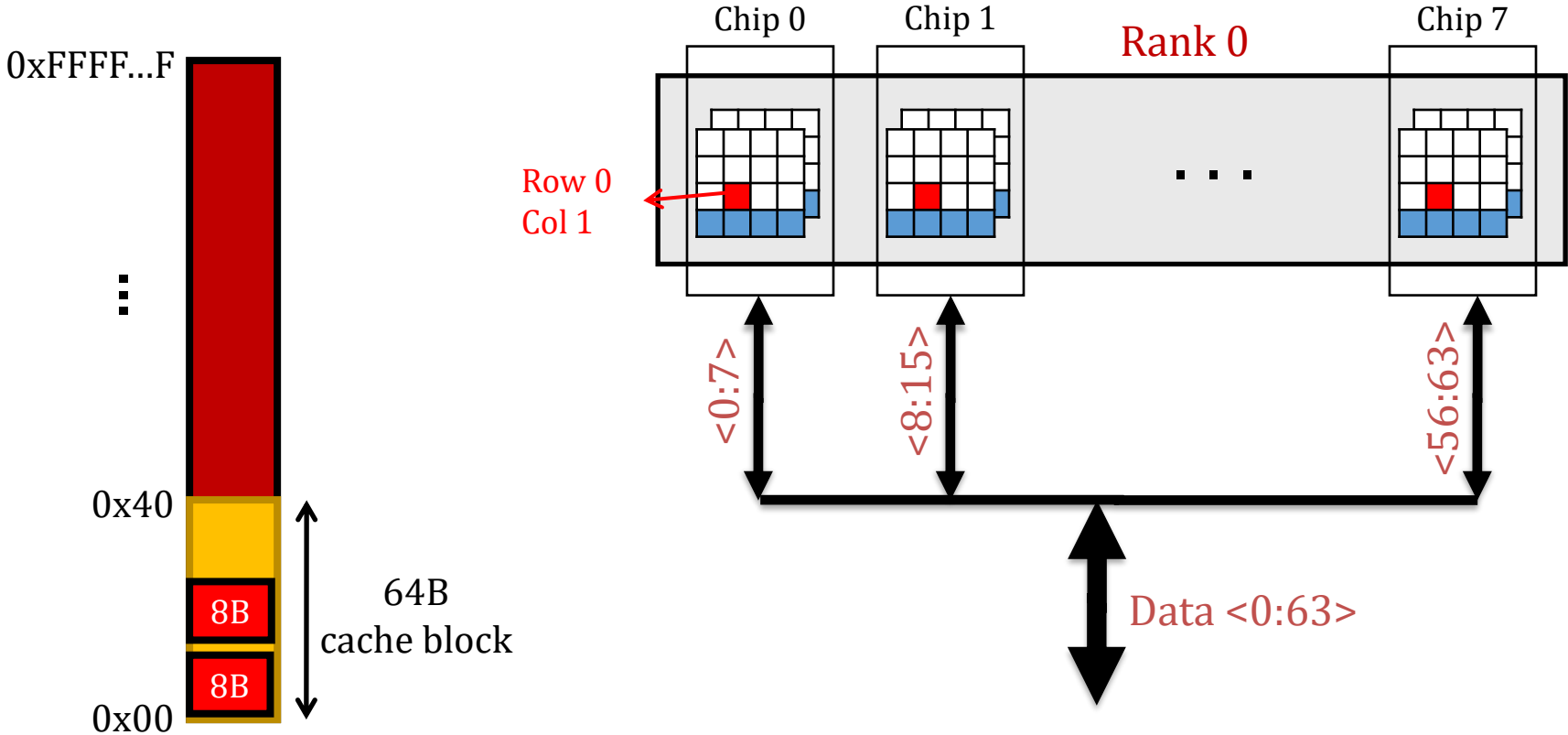
8 cycles (DRAM IO): 1 cycle transfers 8 bytes from a column

DRAM Address Mapping (One Channel)

2GB DRAM, 8 Banks, 16K rows, 2K Columns per bank

Cache Interleaving: Consecutive cache blocks in consecutive banks

Row (14 bits)	High Column	Bank (3 bits)	Low Col.	Byte in bus (3 bits)
	8 bits		3 bits	
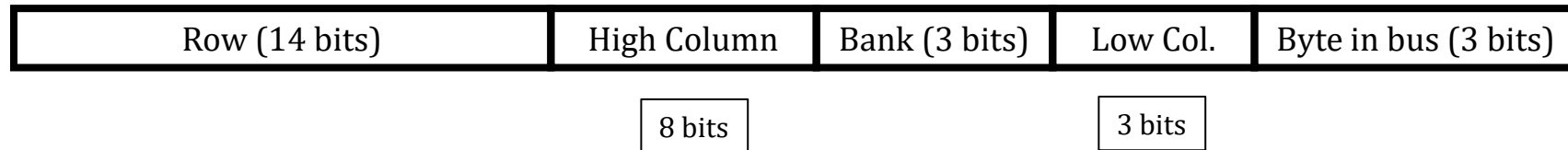
Row Interleaving: Consecutive rows in consecutive banks

Row (14 bits)	Bank (3 bits)	Column (11 bits)	Byte in bus (3 bits)

Row Access

A "DRAM row" is also called a "DRAM page"
"Sense amplifiers" also called "row buffer"

Each address is a <row,column> pair
Access to a "closed row"

- Activate command opens row (placed into row buffer)
- Read/write command reads/writes column in the row buffer
- Precharge command closes the row and prepares the bank for next access

Access to an "open row"

- No need for activate command

Row Buffer hit/miss/conflict

Access Address:
(Row 0, Column 0)
(Row 0, Column 1)
(Row 0, Column 85)
(Row 1, Column 0)

Columns

Row decoder

Rows

Row address 1

Row 1   Row Buffer CONFLICT !

Column address 0

Column mux

Data

Computer Architecture

15
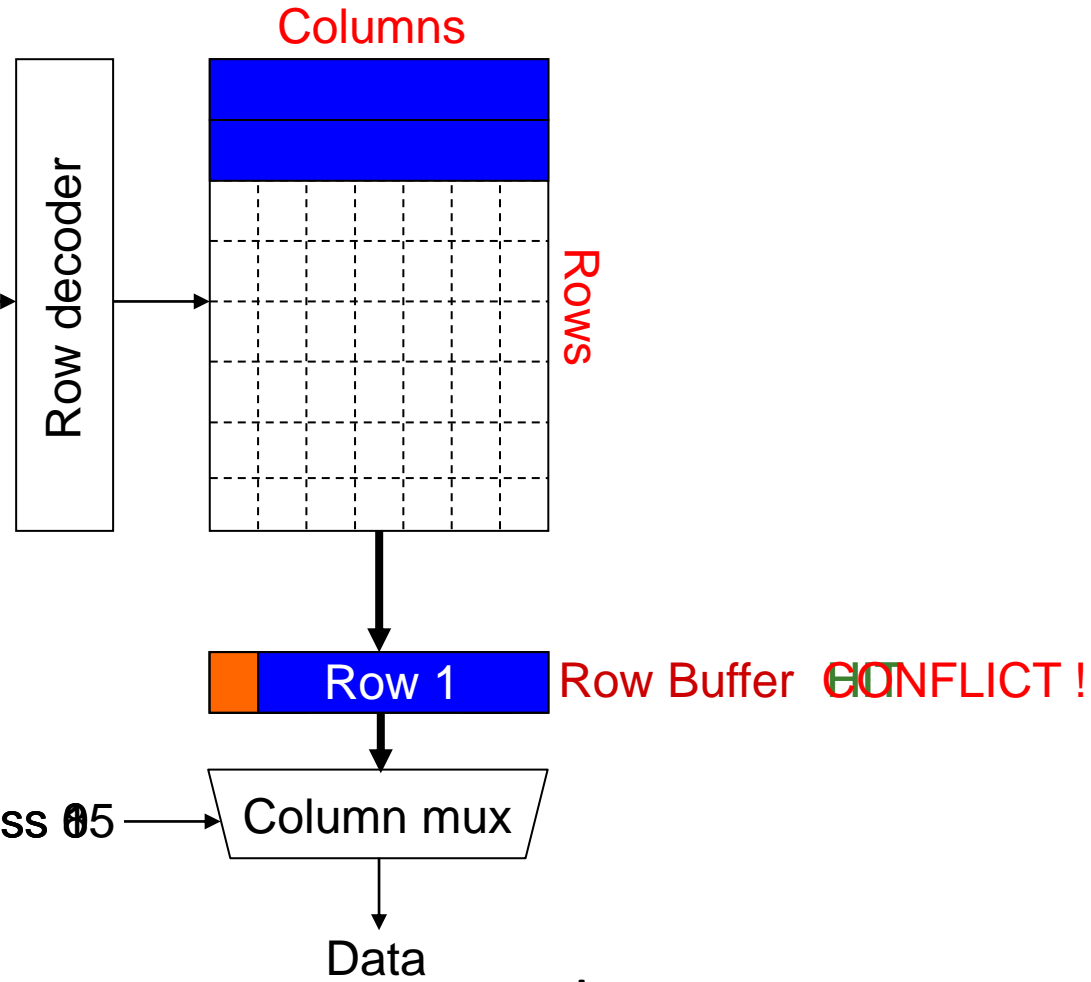
DRAM Timing Constraints

tRAS: Row address strobe latency

tRP: Precharge latency

tCAS: Column address strobe latency

tRC: Row cycle time: tRAS+tRP

Row (Page) Policies

Open Page (do not get confused with an OS page)

After an access: Keep the page in the row-buffer

Consecutive accesses to the same page : Row-buffer Hit

On an access to different page: Close the row and open the new one

Closed page:

After an access: Close the page if there are no accesses to the same row in the request queue.

Consecutive accesses to different page : Low latency

On an access to different page: No need to close the row

10K view on the latency

Page Empty: ACT + CAS

Page Hit: CAS

Page Miss: PRE+ACT+CAS

DRAM Refresh

DRAM cells

Sense Amps

Row Buffer

DRAM cells lose contents after a while,
Refresh command refreshes all rows at different granularity

How to implement a refresh? What is the latency? How frequent?

DRAM cells are refreshed every 64 ms at normal temperature (< 85∘C).

Computer Architecture

DRAM Controller

LLC

DRAM
Controller

Addr. Mapper

RQ | WQ

CQ | CQ | CQ | CQ

Command Scheduler

Command and
Address bus

Double Data Rate (DDR)
Data bus (burst length= eight beats each of 64-bits)

Frequency

800MHz

1066MHz

1333MHz

Frequency

1600MHz

2133MHz

2666MHz

Computer Architecture

Reads and Writes(Writebacks)

Reads are critical to performance

Write Queue stores writes and the writes are serviced
after # writes reach a threshold

Why?

The direction of the data bus changes from reads to writes. So ??

DRAM controller creates DRAM commands from based on the
requests at read Q and write Q

Computer Architecture

DRAM Scheduling

Based on
Row-buffer locality, Source
of the request,
Loads/Stores
Load criticality

Satisfy all the timing constraints. Around 60

FR-FCFS

Prefers requests with Row hits (column-first) FR: First Ready

Names	Memory clock	I/O bus clock	Transfer rate	Theoretical bandwidth
DDR-200, PC-1600	100 MHz	100 MHz	200 MT/s	1.6 GB/s
DDR-400, PC-3200	200 MHz	200 MHz	400 MT/s	3.2 GB/s
DDR2-800, PC2-6400	200 MHz	400 MHz	800 MT/s	6.4 GB/s
DDR3-1600, PC3-12800	200 MHz	800 MHz	1600 MT/s	12.8 GB/s
DDR4-2400, PC4-19200	300 MHz	1200 MHz	2400 MT/s	19.2 GB/s
DDR4-3200, PC4-25600	400 MHz	1600 MHz	3200 MT/s	25.6 GB/s
DDR5-4800, PC5-38400	300 MHz	2400 MHz	4800 MT/s	38.4 GB/s
DDR5-6400, PC5-51200	400 MHz	3200 MHz	6400 MT/s	51.2 GB/s

آپ کا دن اچھا گزرے۔