CASPER

CS305: Computer Architecture

DRAM Timing Constraints

https://www.cse.iitb.ac.in/~biswa/courses/CS305/main.html

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

Row Access

A “DRAM row” is also called a “DRAM page”
“Sense amplifiers” also called “row buffer”

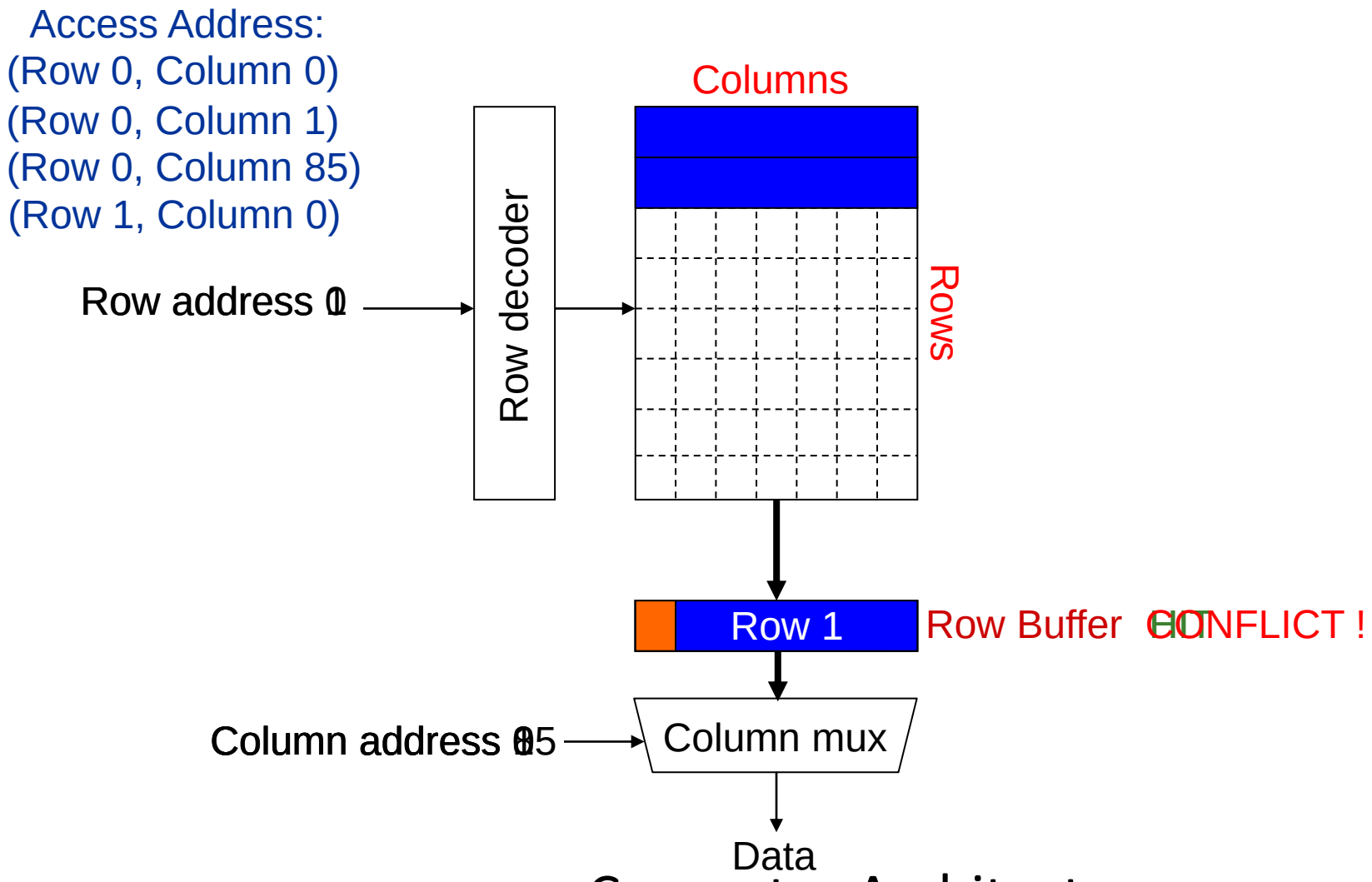
Each address is a <row,column> pair
Access to a “closed row”

- 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



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

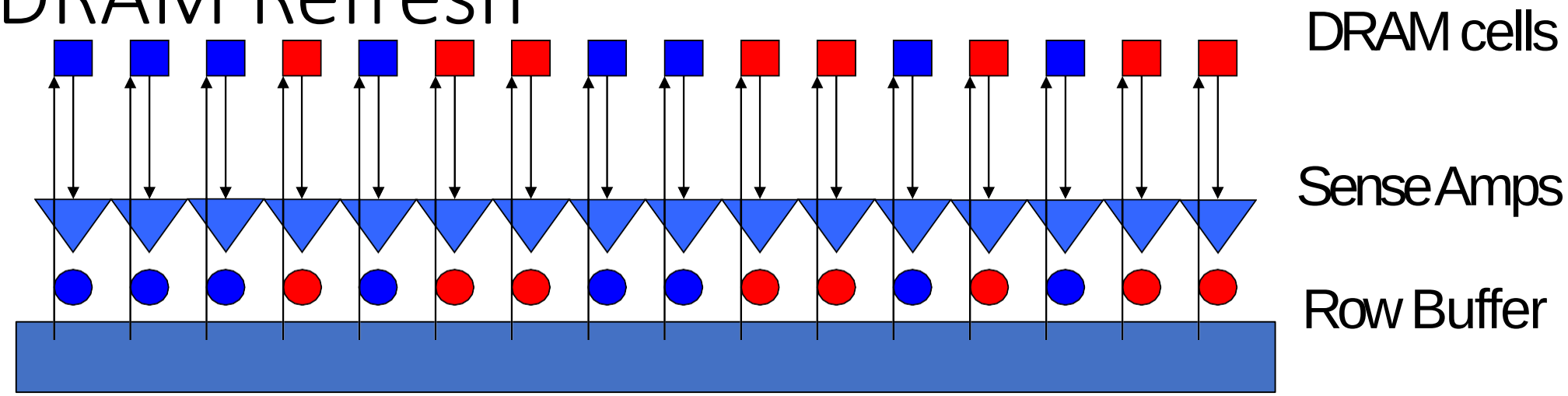
10K view on the latency

Page Empty: ACT + CAS

Page Hit: CAS

Page Miss: PRE+ACT+CAS

DRAM Refresh

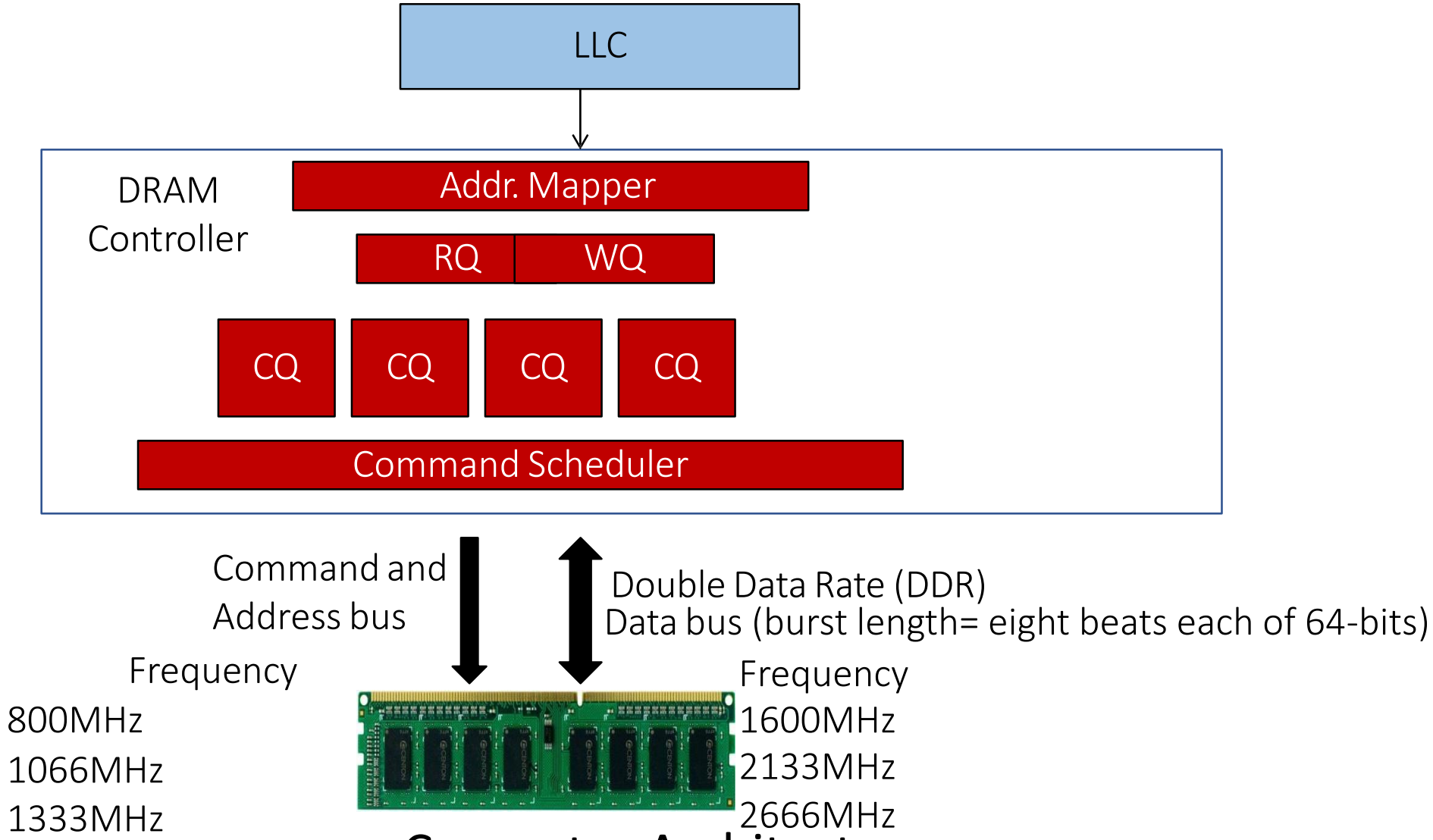


DRAM cells lose contents after awhile,
Refresh command refreshes all rows at different granularity

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

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

DRAM Controller



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

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

DRAM Bandwidth

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

Metrics of Interest: Multicore system with shared resources

Application *i* running on an N-core system

Throughput = $\sum$ IPC (i), does not consider fairness

Individual Slowdown (i) = CPI-together (i) / CPI-alone (i)

Weighted Speedup = $\sum$ (IPC-together(i) / IPC-alone (i))

Hvala