《计算机系统结构》课程直播

2020. 3.12

请将ZOOM名称改为“姓名”；

听不到声音请及时调试声音设备；签到将在课间休息进行

Memories (SRAM & DRAM)

存储器技术与优化

本次讲课:存储技术与优化

1 SRAM和DRAM特点

2 存储器性能优化技术

3 存储系统性能优化技术

SRAM和DRAM的特点

Memory System Architecture

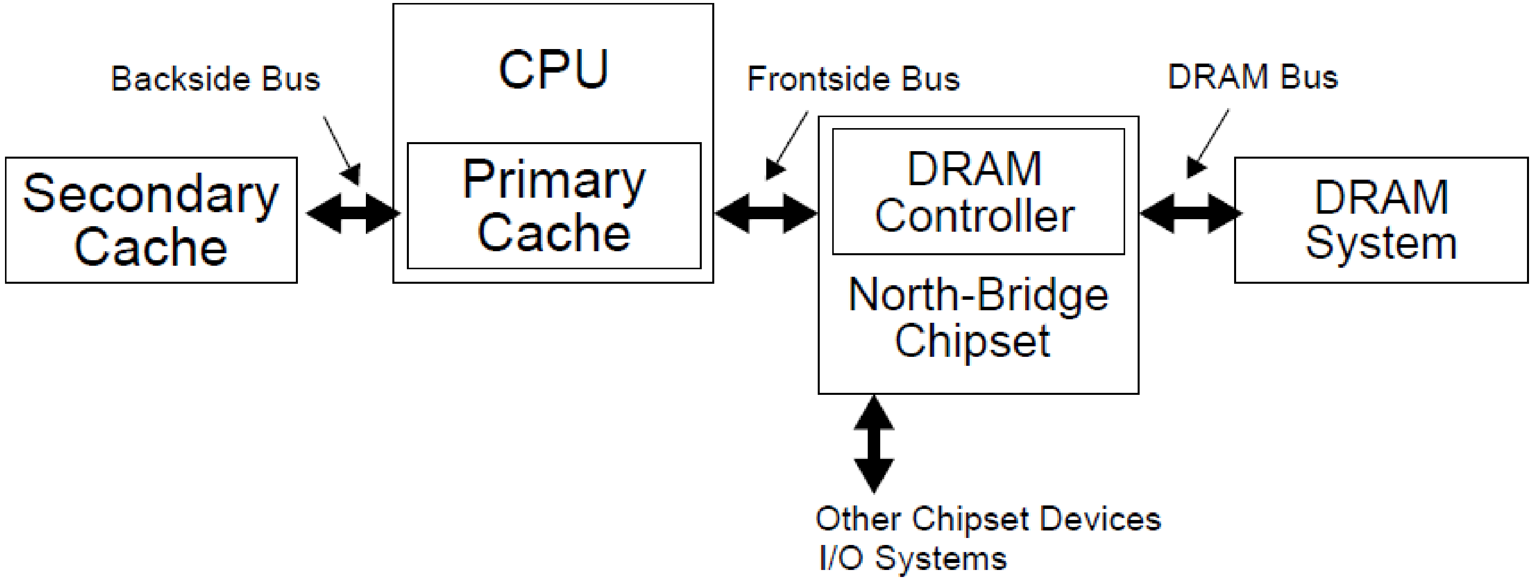


Figure 4.1: Memory System Architecture

参考文献

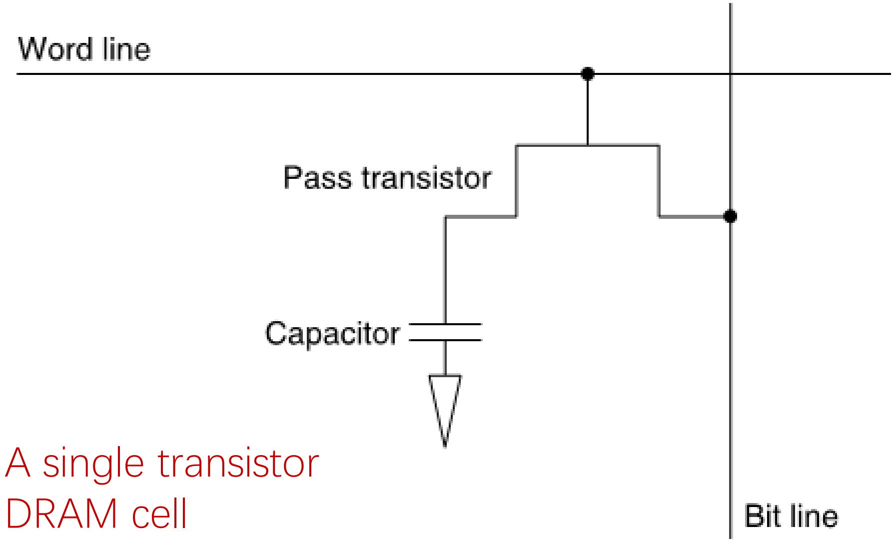
Davis, B. T. (2001). Modern dram architectures, University of Michigan: 221.
Jacob, B. (2009). The Memory System, Morgan & Claypool.

Types of Memory

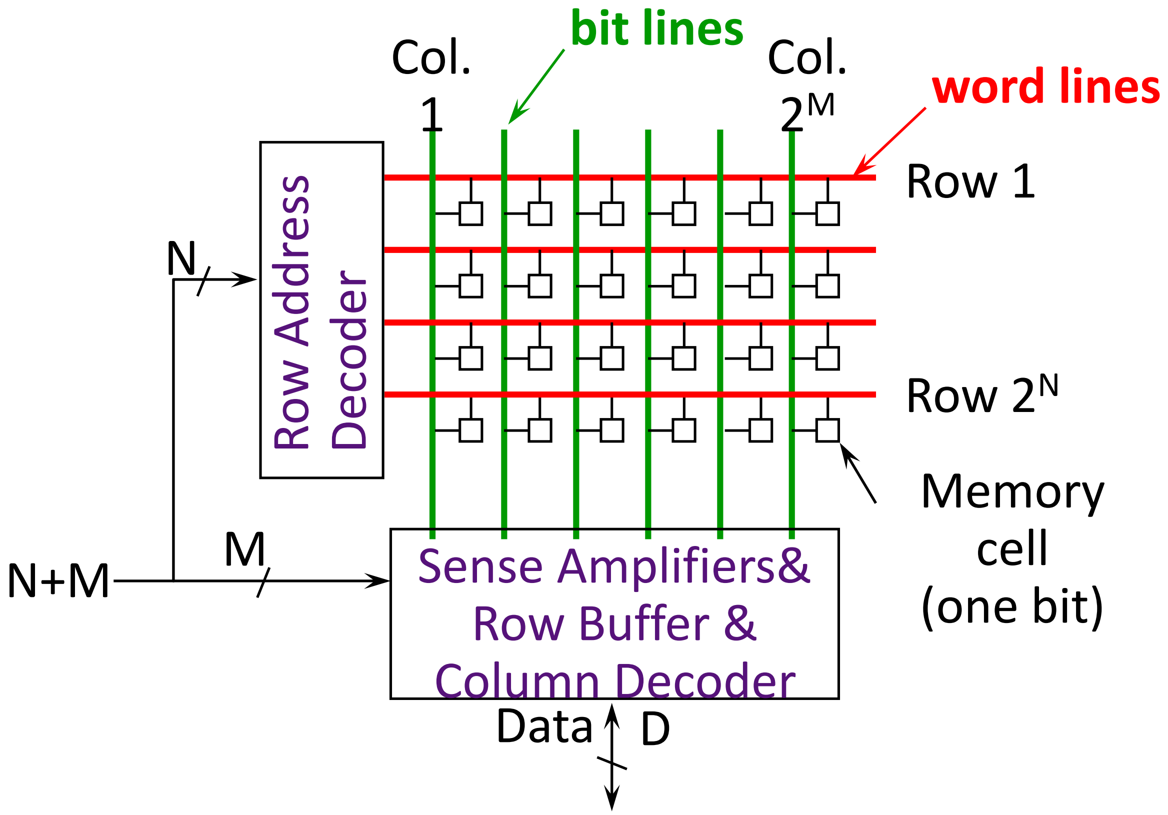
- **Static RAM (SRAM)**
  - Cache: SRAM
  - 6 transistors per bit
    - Two inverters (4 transistors) + transistors for reading/writing
  - Optimized for speed (first) and density (second)
  - Fast (sub-nanosecond latencies for small SRAM)
    - Speed roughly proportional to its area (~ sqrt(number of bits))
- **Dynamic RAM (DRAM)**
  - Memory: DRAM,
  - 1 transistor + 1 capacitor per bit
  - Optimized for density (in terms of cost per bit)
  - Slow (>30ns internal access, ~50ns pin-to-pin)
- Nonvolatile storage: Magnetic disk, Flash RAM, Phase-change memory, …

DRAM

- DRAM
  - 每位1个 transistor
  - 必须要周期性的刷新
  - 地址线复用:
    - Lower half of address: column access strobe (CAS)
    - Upper half of address: row access strobe (RAS)

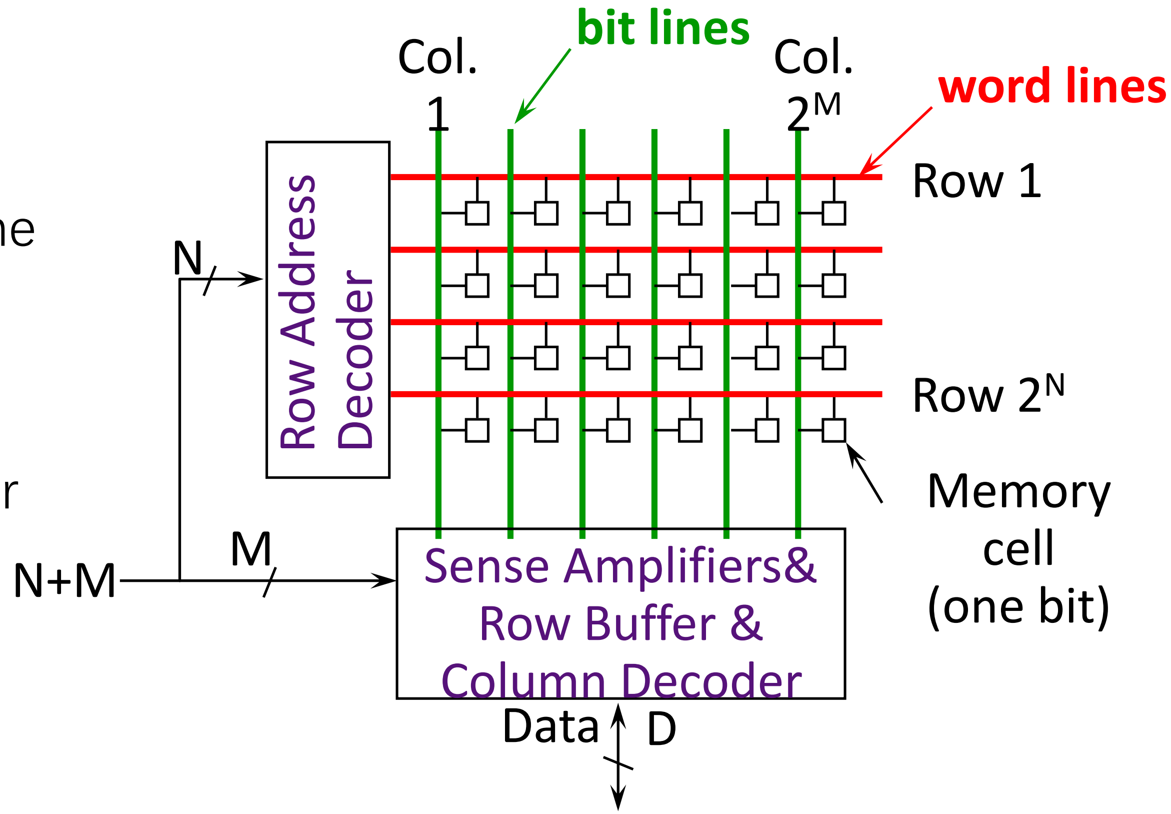


A single transistor DRAM cell



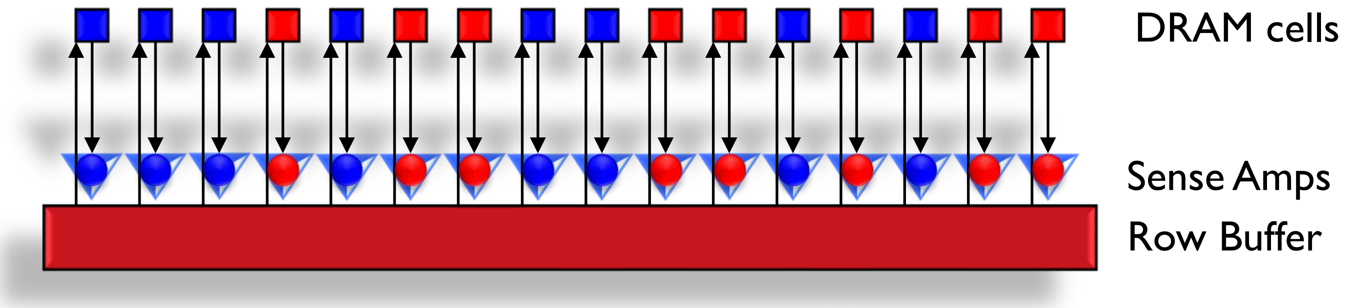
DRAM Chip Organization

- Array organization
- Reads *destructive*: contents are erased by reading
- *Row buffer* holds read data
  - Data in row buffer is called a *DRAM row*
    - Often called "page" - not necessarily same as OS page
  - Read gets entire row into the buffer
  - Reads always performed out of the row buffer
    - Reading a whole row, but accessing one block



DRAM Read

- After a read, the contents of the DRAM cell are gone
  - But still “safe” in the row buffer
- Write bits back before doing another read
- Reading into buffer is slow, but reading buffer is fast
  - Try reading multiple lines from buffer *(row-buffer hit)*

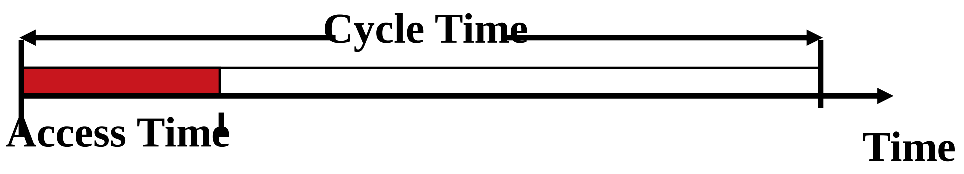


Process is called *opening* or *closing* a row

存储器性能优化

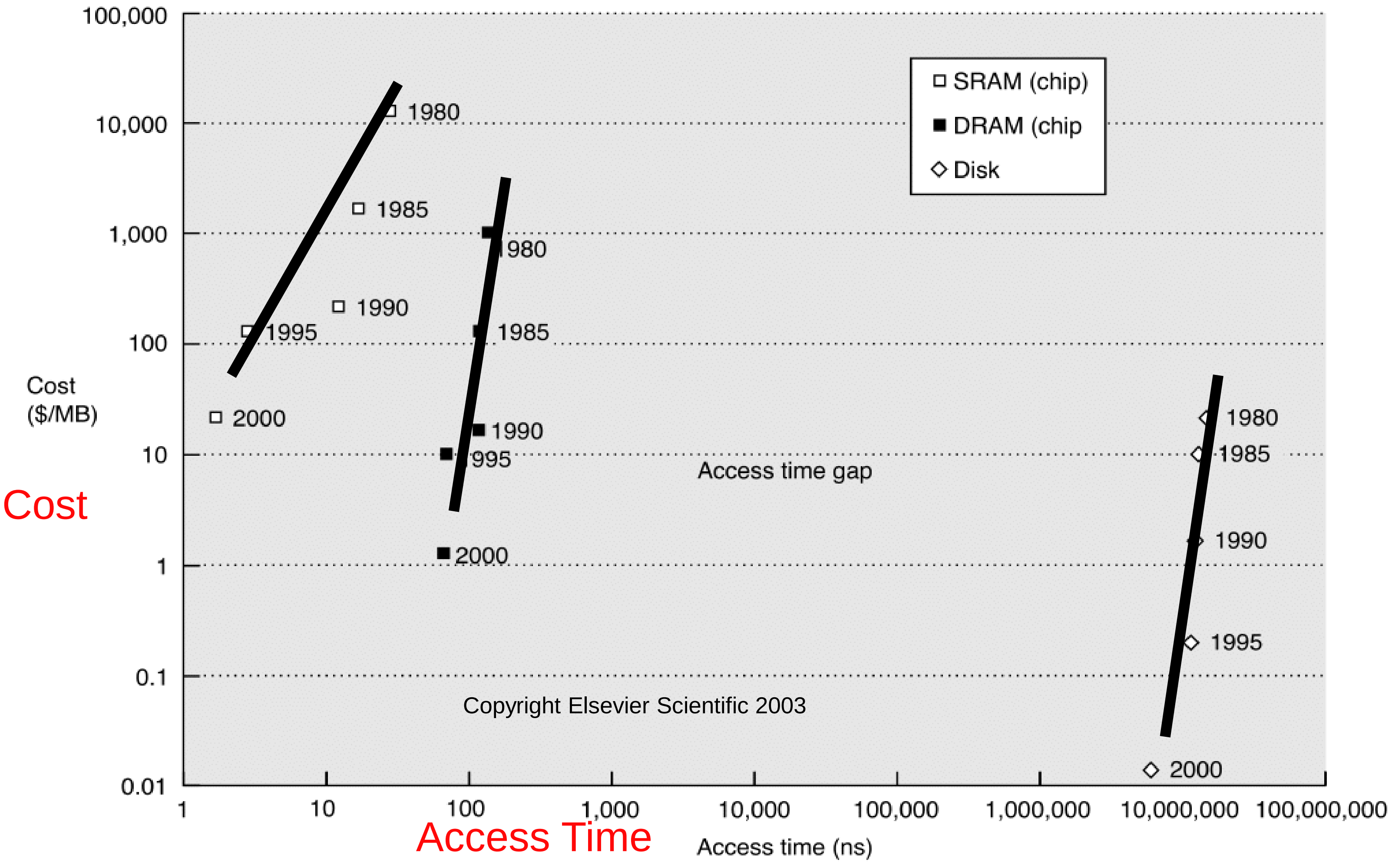
存储器技术

- 存储器的访问
  - 取指令、取操作数、写操作数和I/O
- 存储器性能指标
  - 容量、速度和每位价格
  - 访问时间/访存延迟（Access Time /Latency)
  - 存储周期（Cycle Time)



  - 存储器带宽（Bandwidth)

Memory Technology Trends

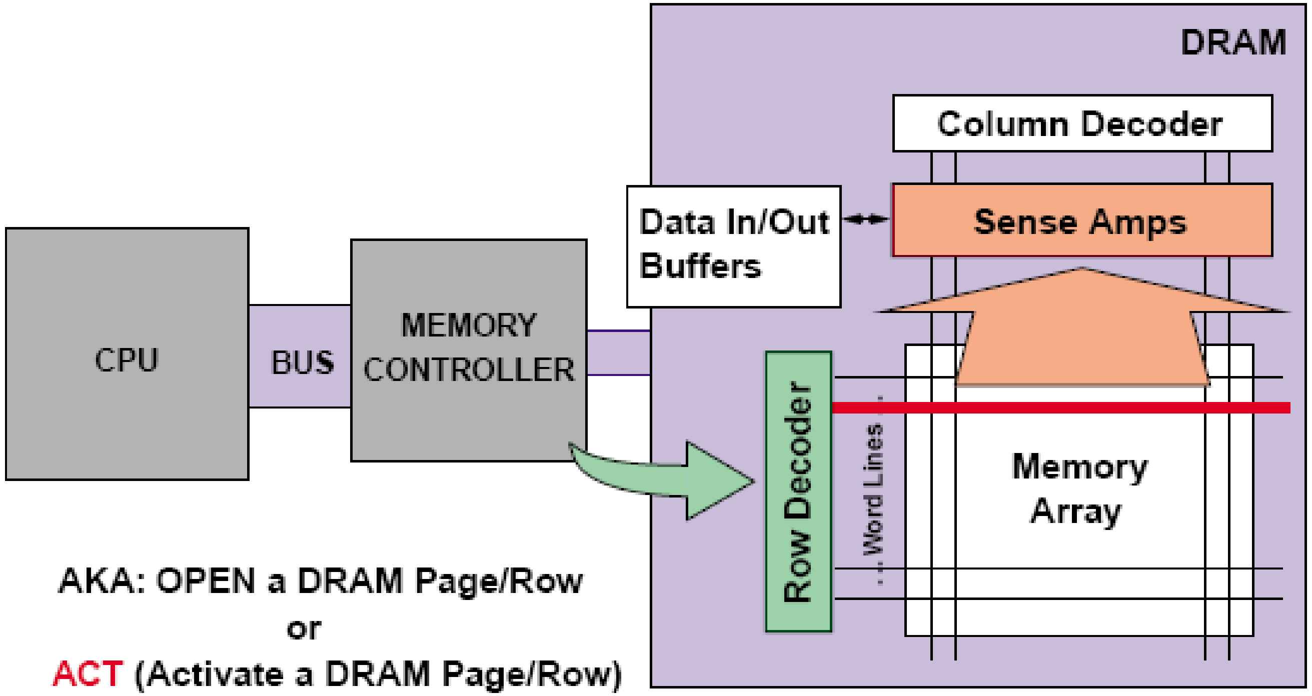


Cost

Access Time

Typical DRAM Access Sequence (1/5)

[PRECHARGE and] ROW ACCESS



AKA: OPEN a DRAM Page/Row

or

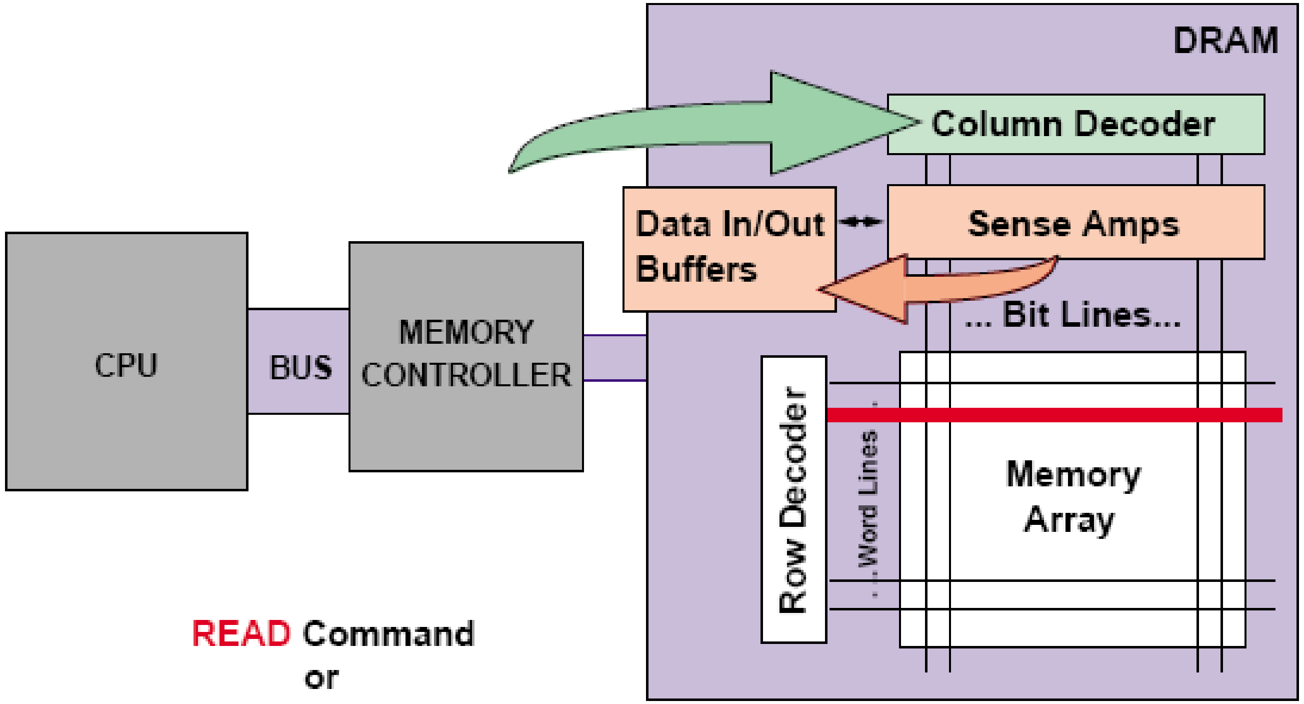
ACT (Activate a DRAM Page/Row)

or

RAS (Row Address Strobe)

Typical DRAM Access Sequence (2/5)

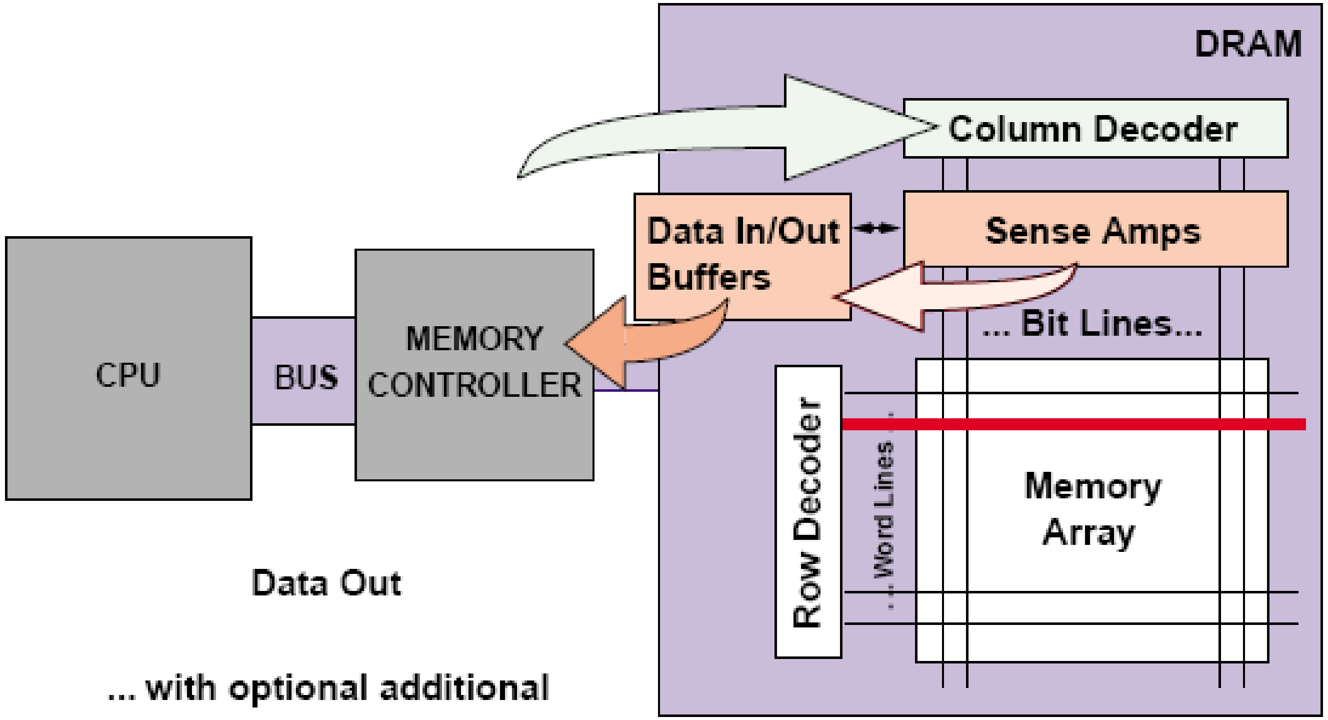
COLUMN ACCESS



READ Command
or
CAS: Column Address Strobe

Typical DRAM Access Sequence (3/5)

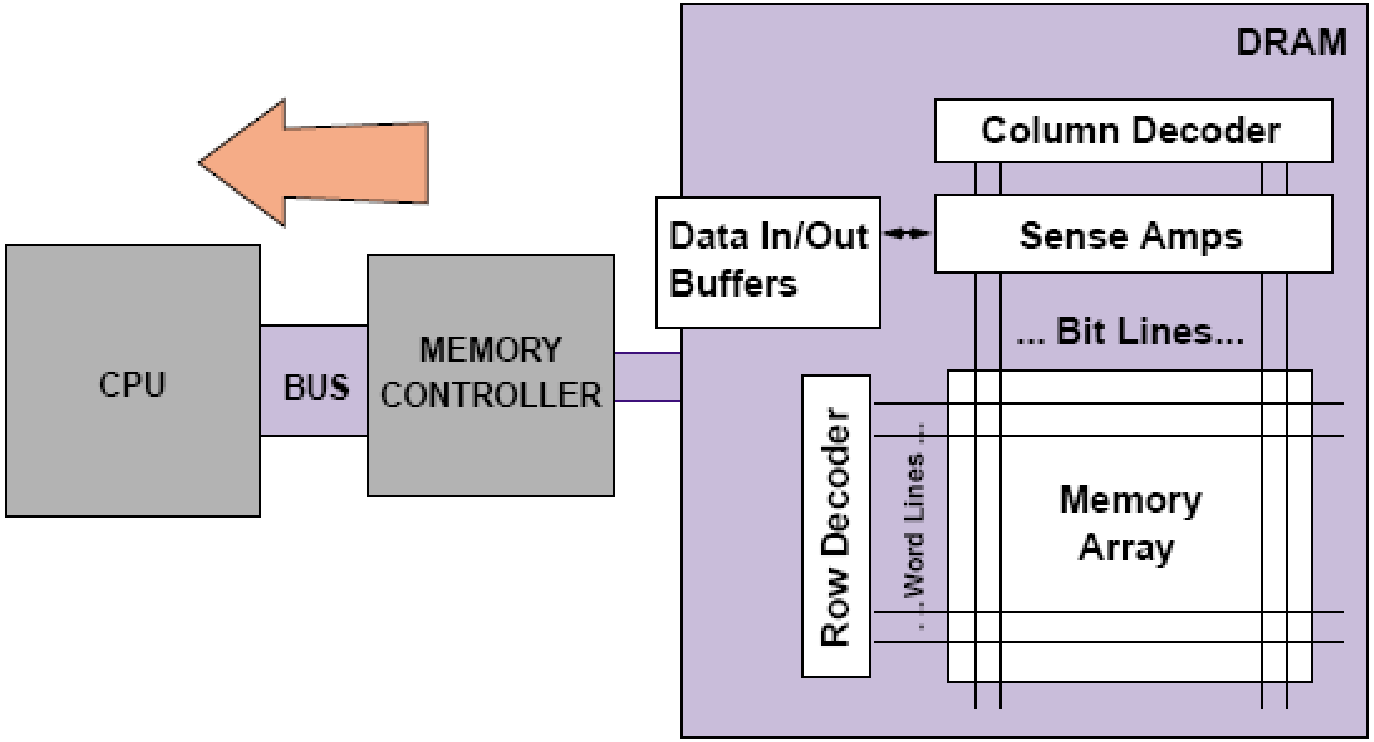
DATA TRANSFER



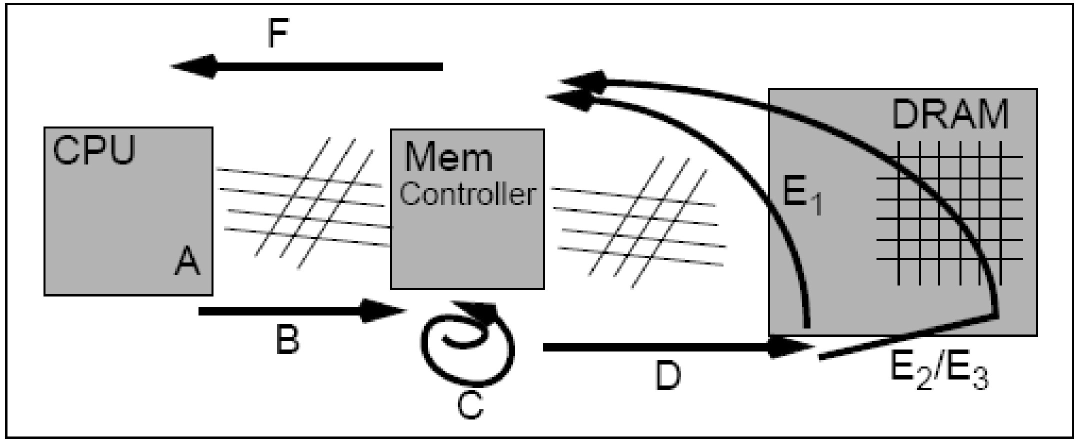
... with optional additional
CAS: Column Address Strobe

Typical DRAM Access Sequence (4/5)

BUS TRANSMISSION



Typical DRAM Access Sequence (5/5)

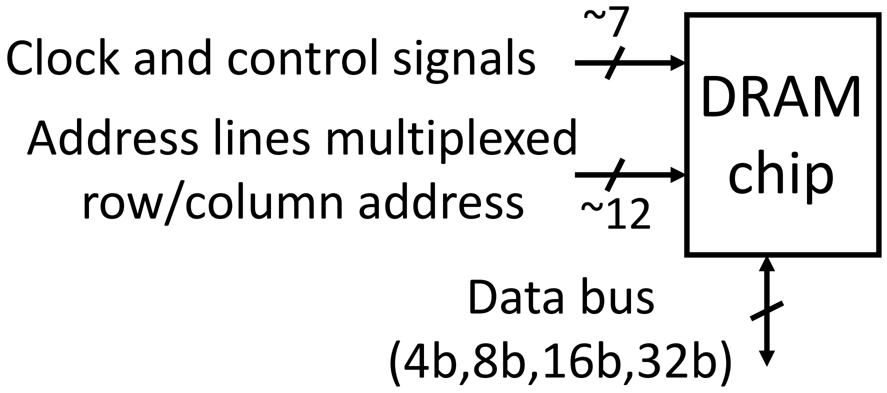


A: Transaction request may be delayed in Queue
B: Transaction request sent to Memory Controller
C: Transaction converted to Command Sequences
(may be queued)
D: Command/s Sent to DRAM
E_1: Requires only a **CAS** or
E_2: Requires **RAS + CAS** or
E_3: Requires **PRE + RAS + CAS**
F: Transaction sent back to CPU

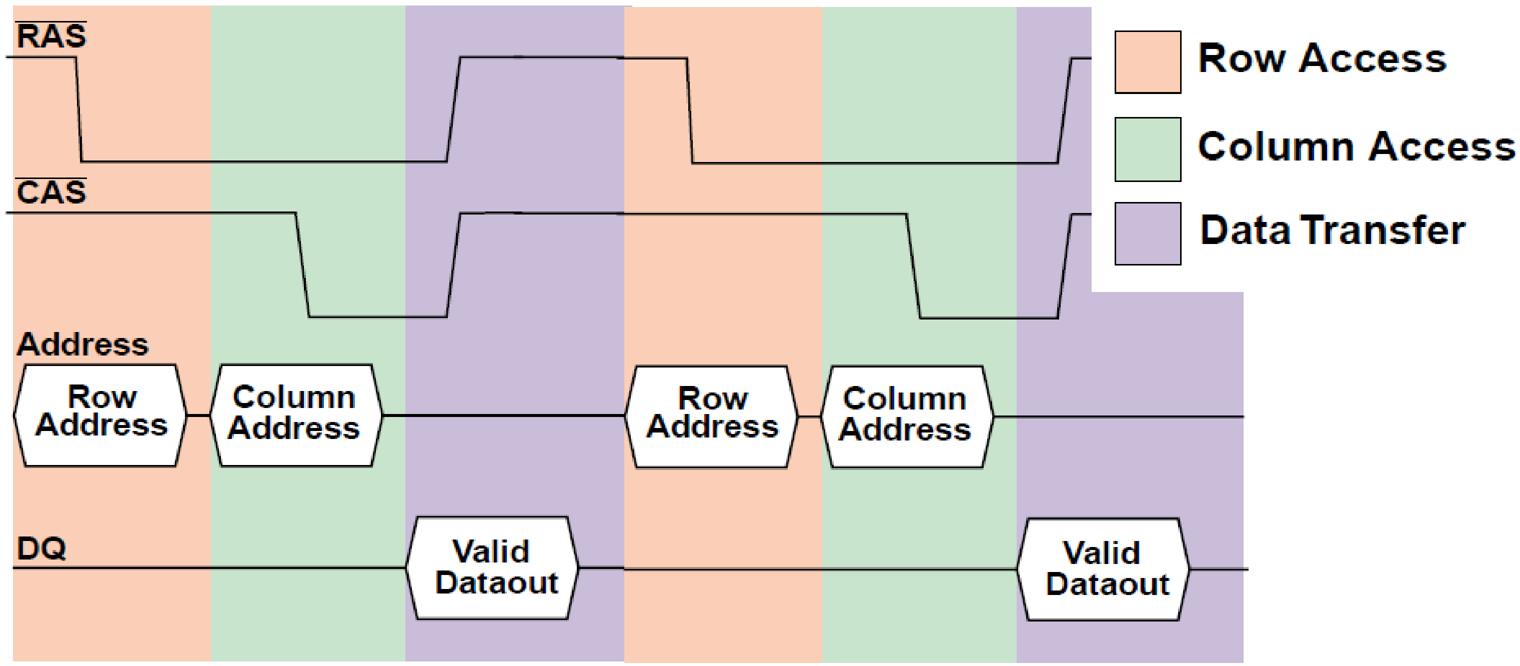
"DRAM Latency" = A + B + C + D + E + F

Performance optimizations

- 如何降低存储器芯片的平均访存延迟、增加带宽?
- Some optimizations:
  - Fast Page Mode Operation
    - Multiple accesses to same row
  - Synchronous DRAM
    - Added clock to DRAM interface
    - Burst mode with critical word first
  - Double data rate (DDR)
  - Wider interfaces
  - Multiple banks on each DRAM device

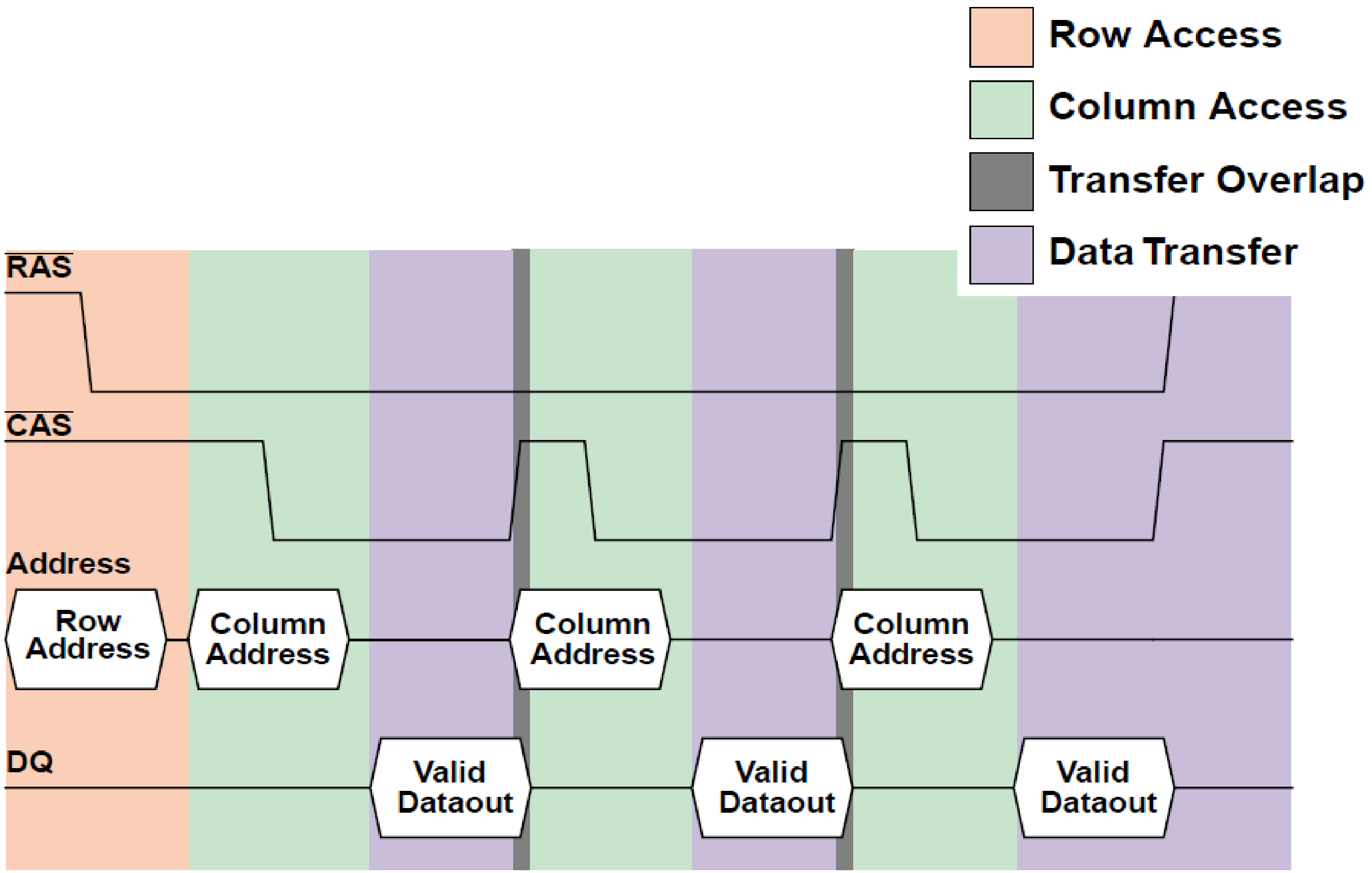


DRAM Read Timing



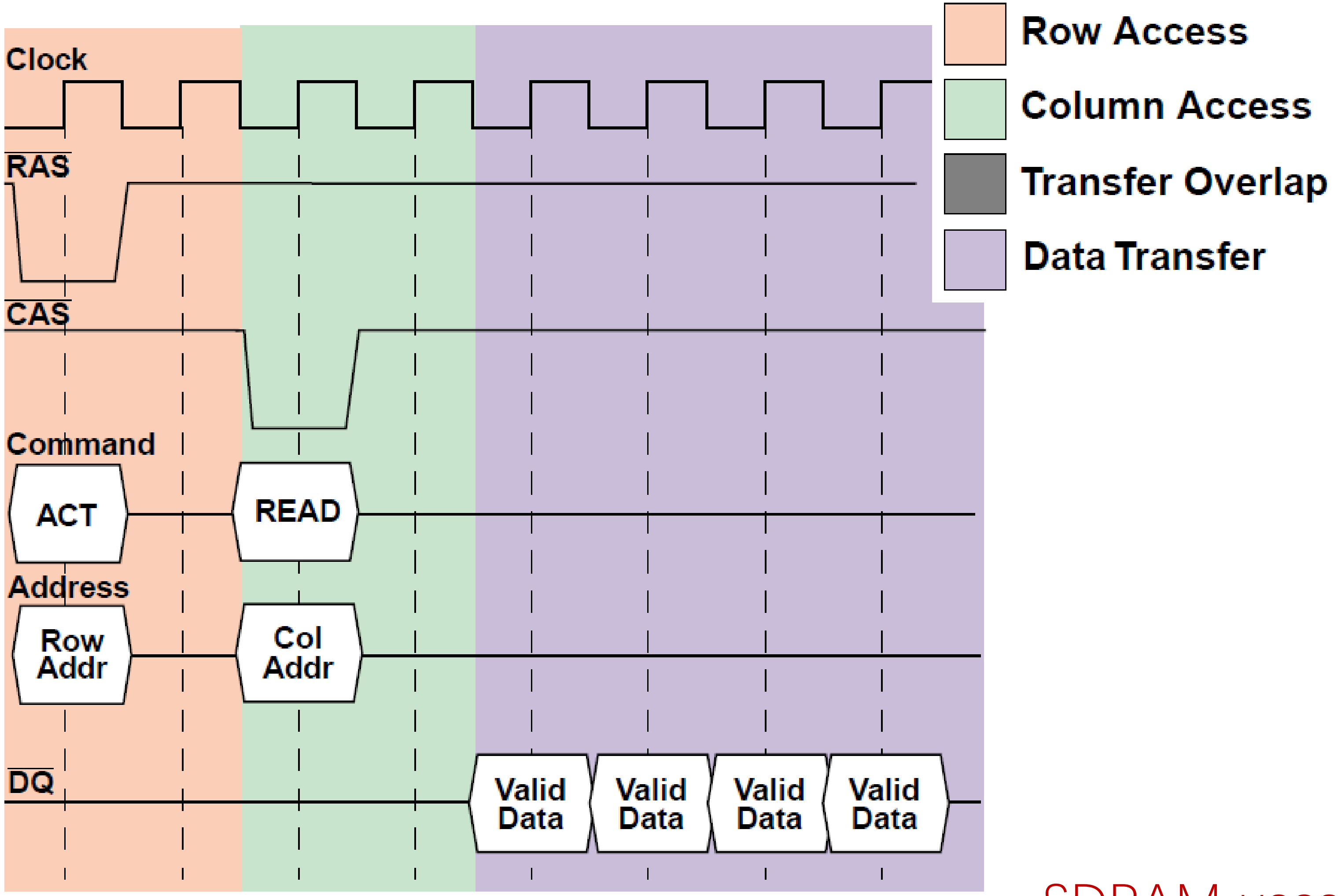
Original DRAM specified Row & Column every time

DRAM Read Timing with Fast-Page Mode



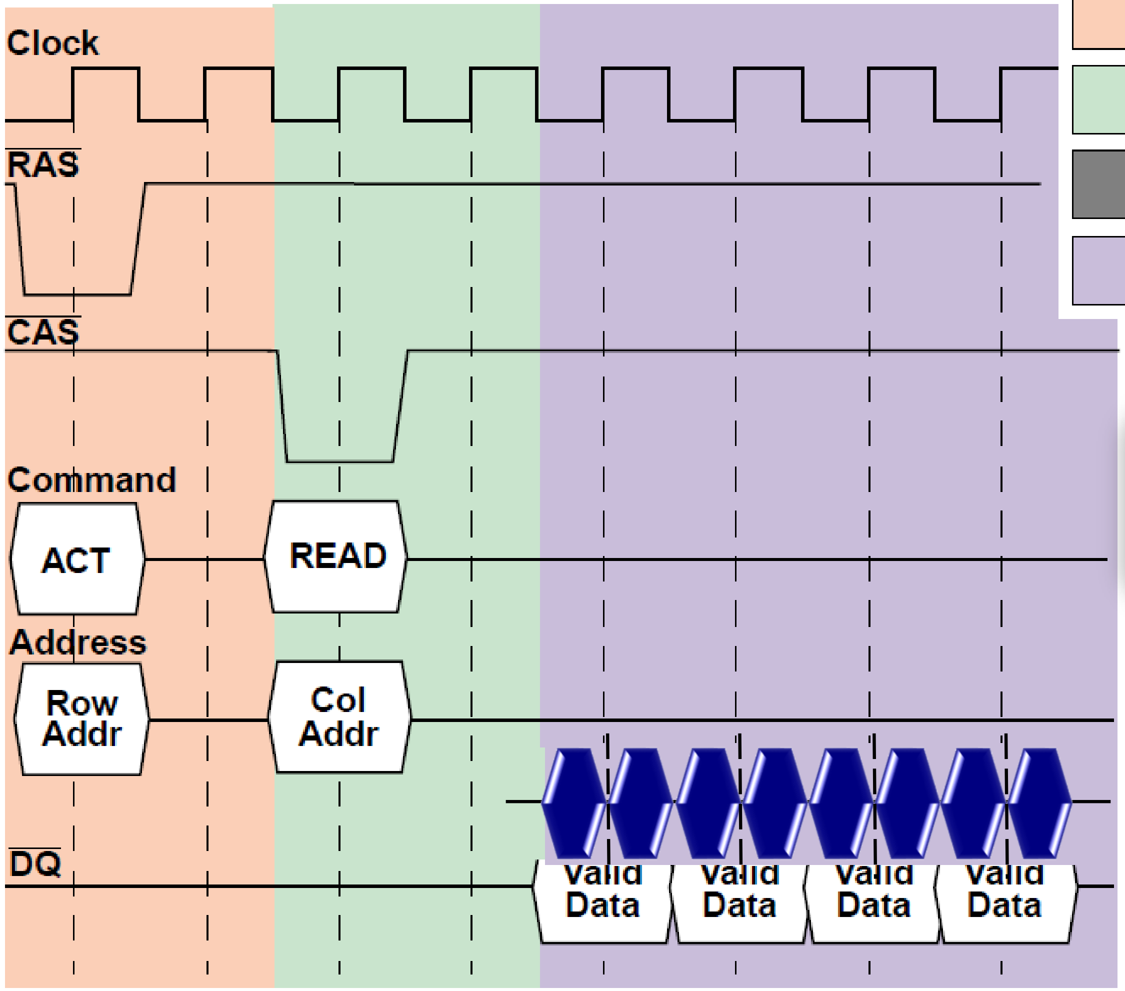
FPM enables multiple reads from page without RAS

SDRAM Read Timing



SDRAM uses clock, supports bursts

DDR-SDRAM Read Timing

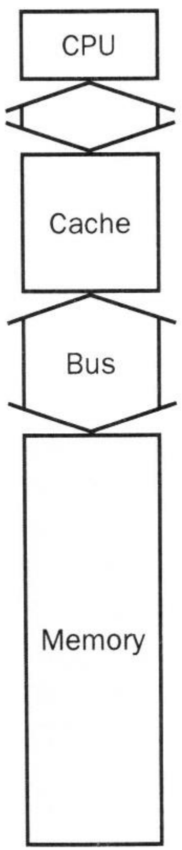


Row Access
Column Access
Transfer Overlap
Data Transfer

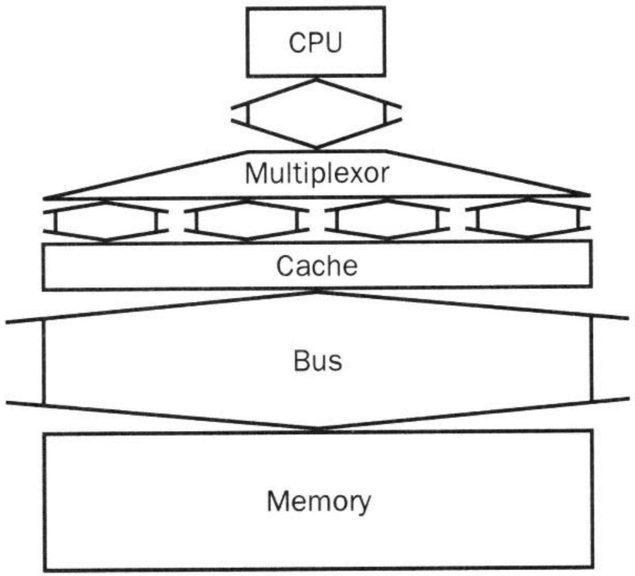
Double-Data Rate (DDR) DRAM transfers data on **both** rising and falling edge of the clock

SDRAM uses clock, supports bursts

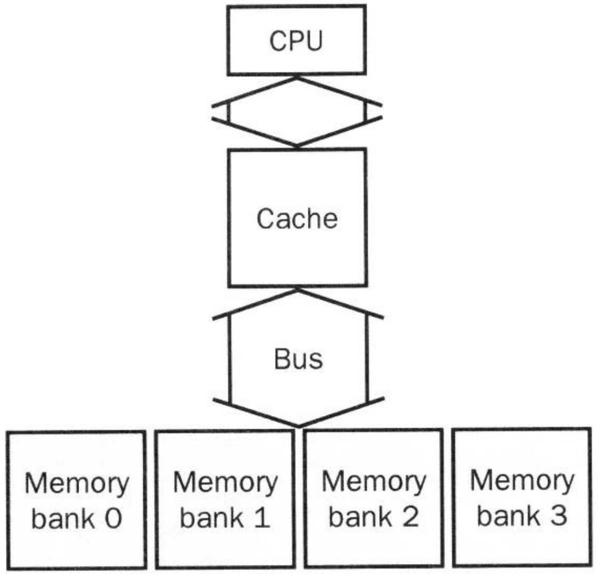
三种存储器组织方式



a. One-word-wide memory organization



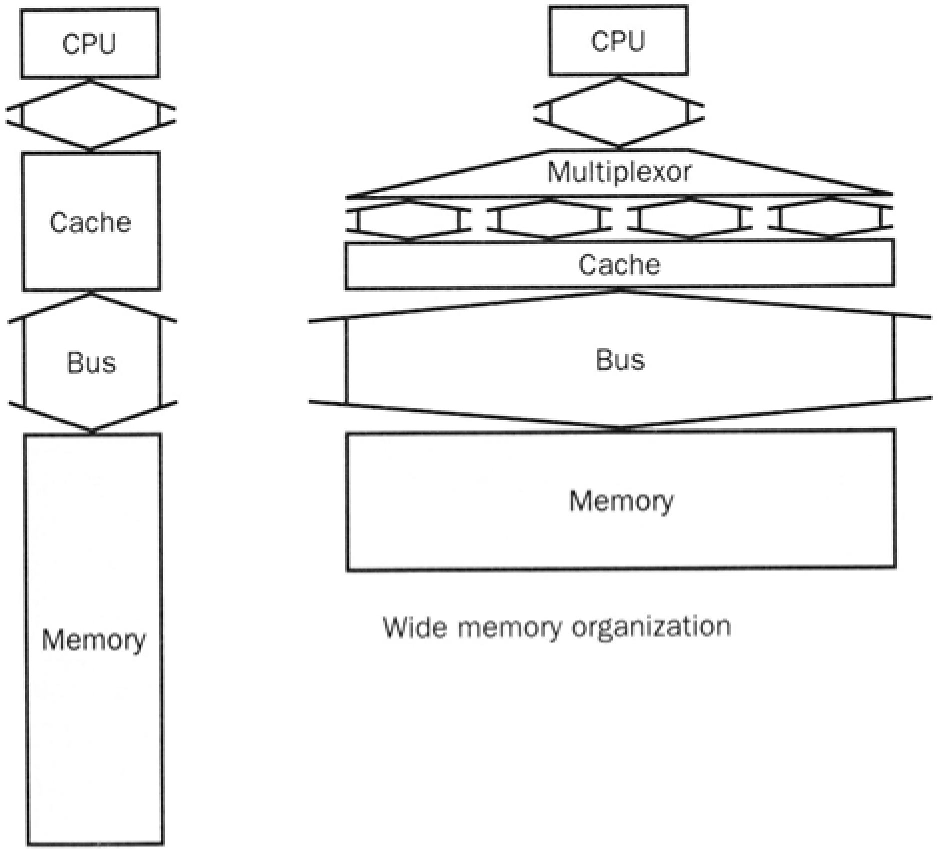
b. Wide memory organization



c. Interleaved memory organization

Wide Bus: Performance

- Miss penalty for an 8-word cache block
  - 1 cycle to send address （cycle: 存储总线周期）
  - 6 cycles to access each word
  - 1 cycle to send word back
  - (1 + 6 + 1) x 8 = 64
- (Expensive) Wider bus option
  - Read all words in parallel
- Miss penalty for 8-word block: 1 + 6 + 1 = 8



Wide memory organization

增大存储器的宽度（并行访问存储器）

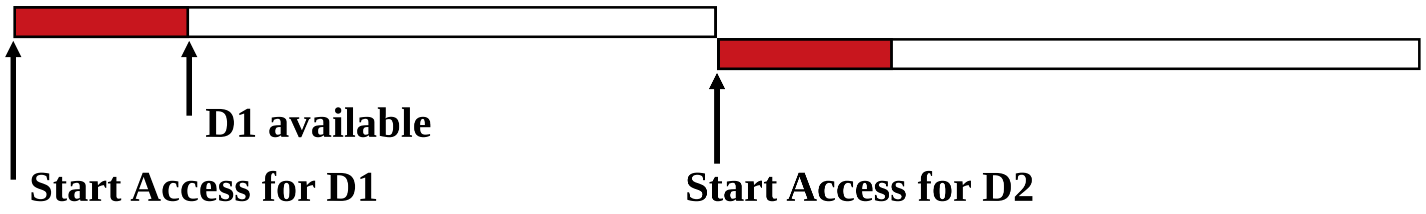
- 最简单直接的方法
- 优点：简单、直接，可有效增加带宽
- 缺点
  - 增加了CPU与存储器之间的连接通路的宽度，实现代价提高
  - 主存容量扩充时，增量应该是存储器的宽度
- 冲突问题
  - 取指令冲突，遇到程序转移时，一个存储周期中读出的n条指令中，后面的指令将无用
  - 读操作数冲突。一次同时读出的几个操作数，不一定都有用
  - 写操作冲突。这种并行访问，必须凑齐n个字之后一起写入。如果只写一个字，必须先把属于同一个存储字的数据读到数据寄存器中，然后在地址码的控制下修改其中一个字，最后一起写。
  - 读写冲突。当要读写的字在同一个存储字内时，无法并行操作。

采用简单的多体交叉存储器

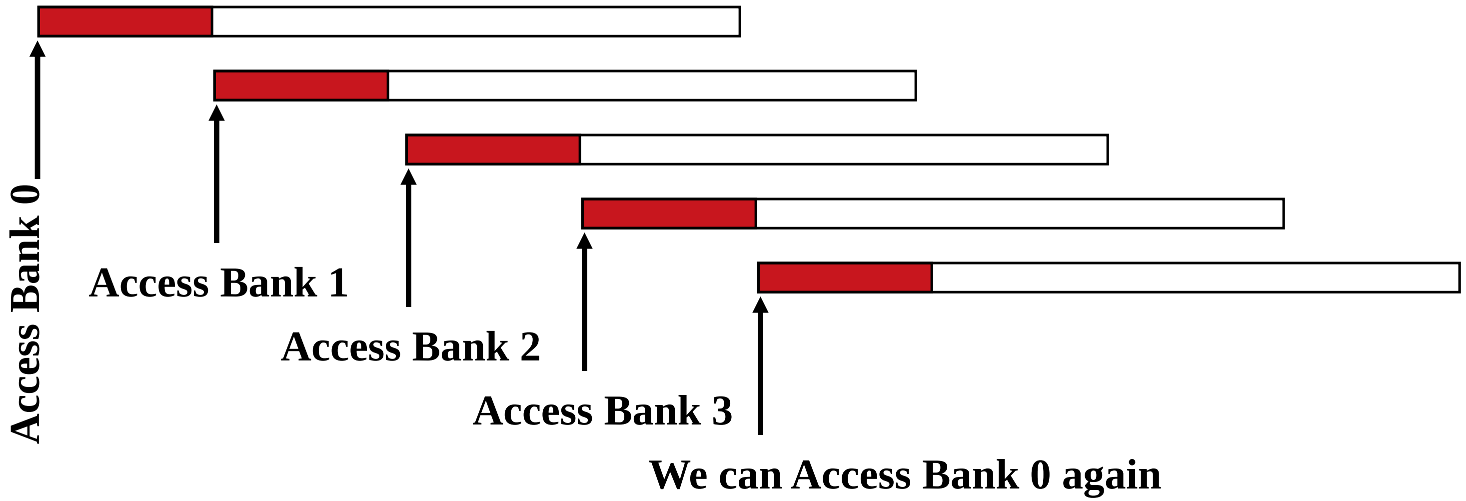
- 一套地址寄存器和控制逻辑
- 存储体的宽度，通常为一个字，不需要改变总线的宽度
- 目的：在总线宽度不变的情况下，完成多个字的并行读写
- 存储器组织为多个体（Bank）
- Divide memory into n banks：“interleave” addresses across them
- Access one bank while another is busy
- Use parallelism in memory banks to hide latency
- 存储模块中所包含的体数，为避免访问冲突，基本原则为：
  - 体的数目 >= 访问体中一个字所需的时钟周期数
- 缺陷：不能对单个体单独访问，对解决冲突没有帮助，逻辑上是一种宽存储器，对各个存储体的访问被安排在不同的时间段

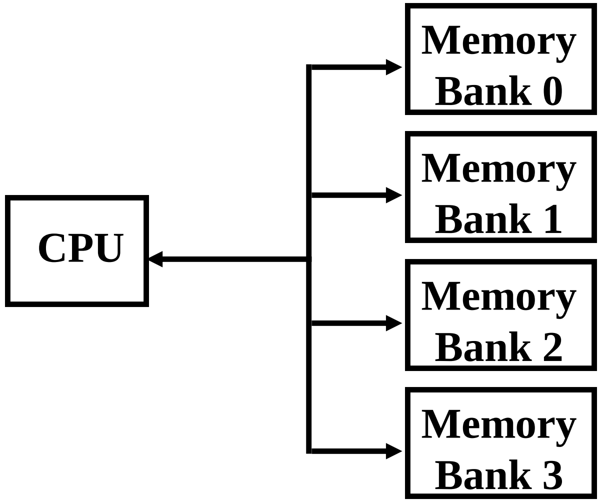
Increasing Bandwidth - Interleaving

Access Pattern without Interleaving:



Access Pattern with 4-way Interleaving:





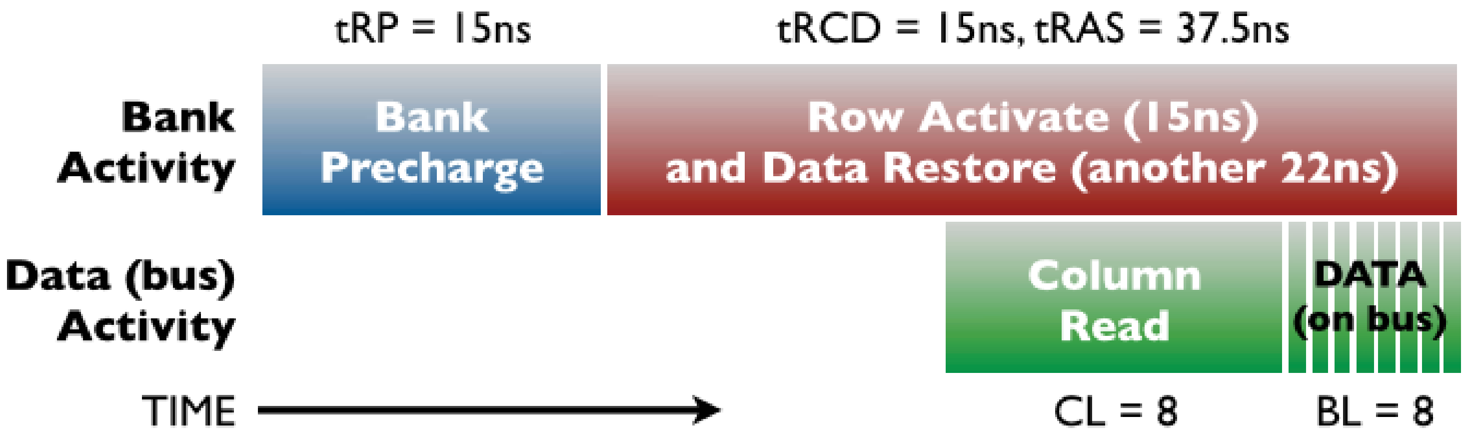
Access time for DDR SDRAM

			Best case access time (no precharge)			Precharge needed
Production year	Chip size	DRAM type	RAS time (ns)	CAS time (ns)	Total (ns)	Total (ns)
2000	256M bit	DDR1	21	21	42	63
2002	512M bit	DDR1	15	15	30	45
2004	1G bit	DDR2	15	15	30	45
2006	2G bit	DDR2	10	10	20	30
2010	4G bit	DDR3	13	13	26	39
2016	8G bit	DDR4	13	13	26	39

Figure 2.4 Capacity and access times for DDR SDRAMs by year of production. Access time is for a random memory word and assumes a new row must be opened. If the row is in a different bank, we assume the bank is precharged; if the row is not open, then a precharge is required, and the access time is longer. As the number of banks has increased, the ability to hide the precharge time has also increased. DDR4 SDRAMs were initially expected in 2014, but did not begin production until early 2016.

From: Computer Architecture A Quantitative Approach (6th Edition)

Cost of Accessing DRAM



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 write back, then read new row, then move data

Bandwidth of DDR SDRAM

Standard	I/O clock rate	M transfers/s	DRAM name	MiB/s/DIMM	DIMM name
DDR1	133	266	DDR266	2128	PC2100
DDR1	150	300	DDR300	2400	PC2400
DDR1	200	400	DDR400	3200	PC3200
DDR2	266	533	DDR2-533	4264	PC4300
DDR2	333	667	DDR2-667	5336	PC5300
DDR2	400	800	DDR2-800	6400	PC6400
DDR3	533	1066	DDR3-1066	8528	PC8500
DDR3	666	1333	DDR3-1333	10,664	PC10700
DDR3	800	1600	DDR3-1600	12,800	PC12800
DDR4	1333	2666	DDR4-2666	21,300	PC21300

Figure 2.5 Clock rates, bandwidth, and names of DDR DRAMS and DIMMs in 2016. Note the numerical relationship between the columns. The third column is twice the second, and the fourth uses the number from the third column in the name of the DRAM chip. The fifth column is eight times the third column, and a rounded version of this number is used in the name of the DIMM. DDR4 saw significant first use in 2016.

Names of DDR SDRAM

- DDR:
  - DDR2： Lower power (2.5 V -> 1.8 V)， Higher clock rates (266 MHz, 333 MHz, 400 MHz)
  - DDR3： 1.5 V， 800 MHz
  - DDR4： 1-1.2 V， 1600 MHz
- GDDR5 is graphics memory based on DDR3
- Graphics memory:
  - Achieve 2-5 X bandwidth per DRAM vs. DDR3
    - Wider interfaces (32 vs. 16 bit)
    - Higher clock rate

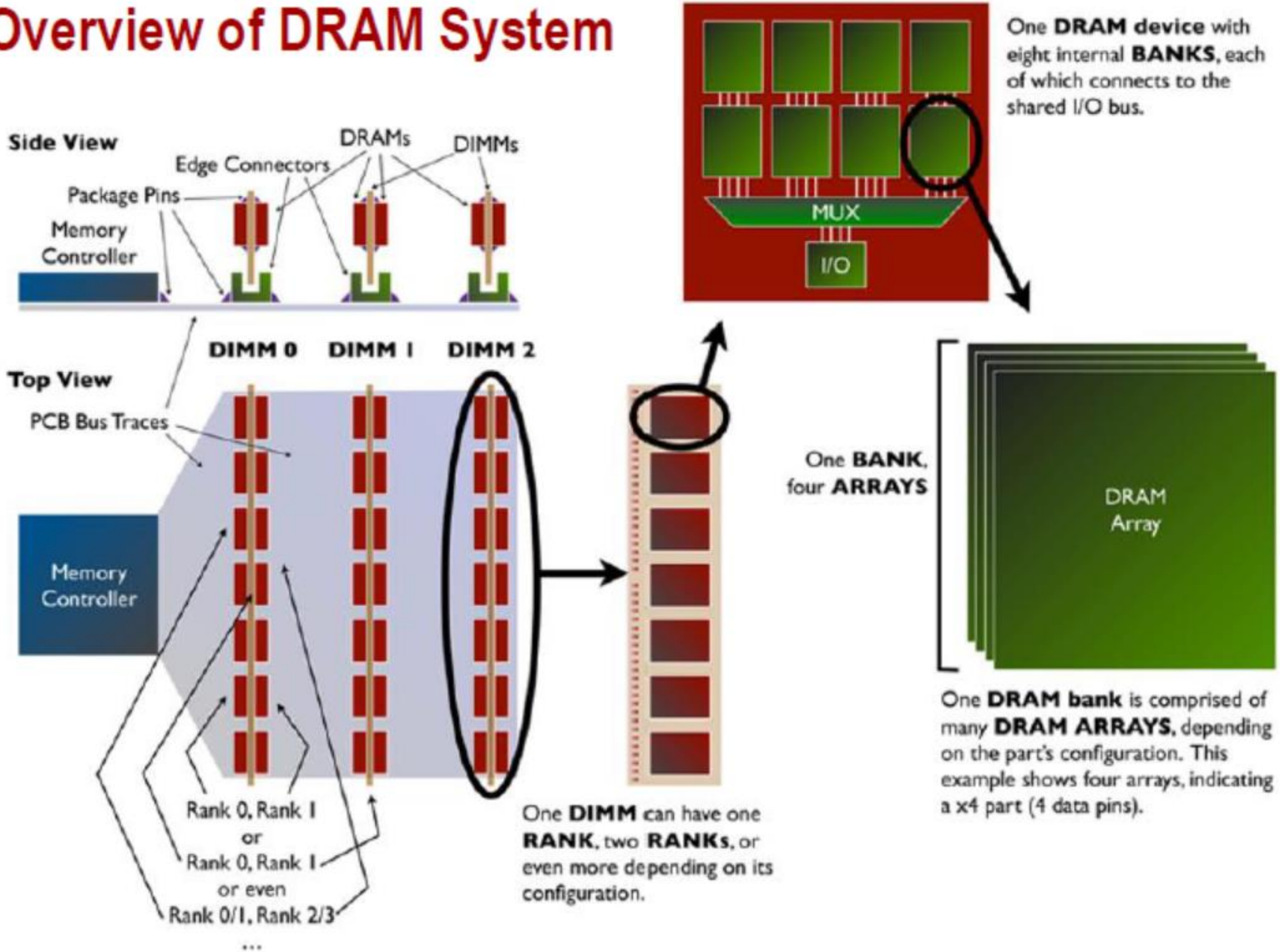
练习

DDR3 SDRAM 芯片内部核心频率是133.25Mhz, 与之相连的存储总线每次传输8B，下面描述错误的是:

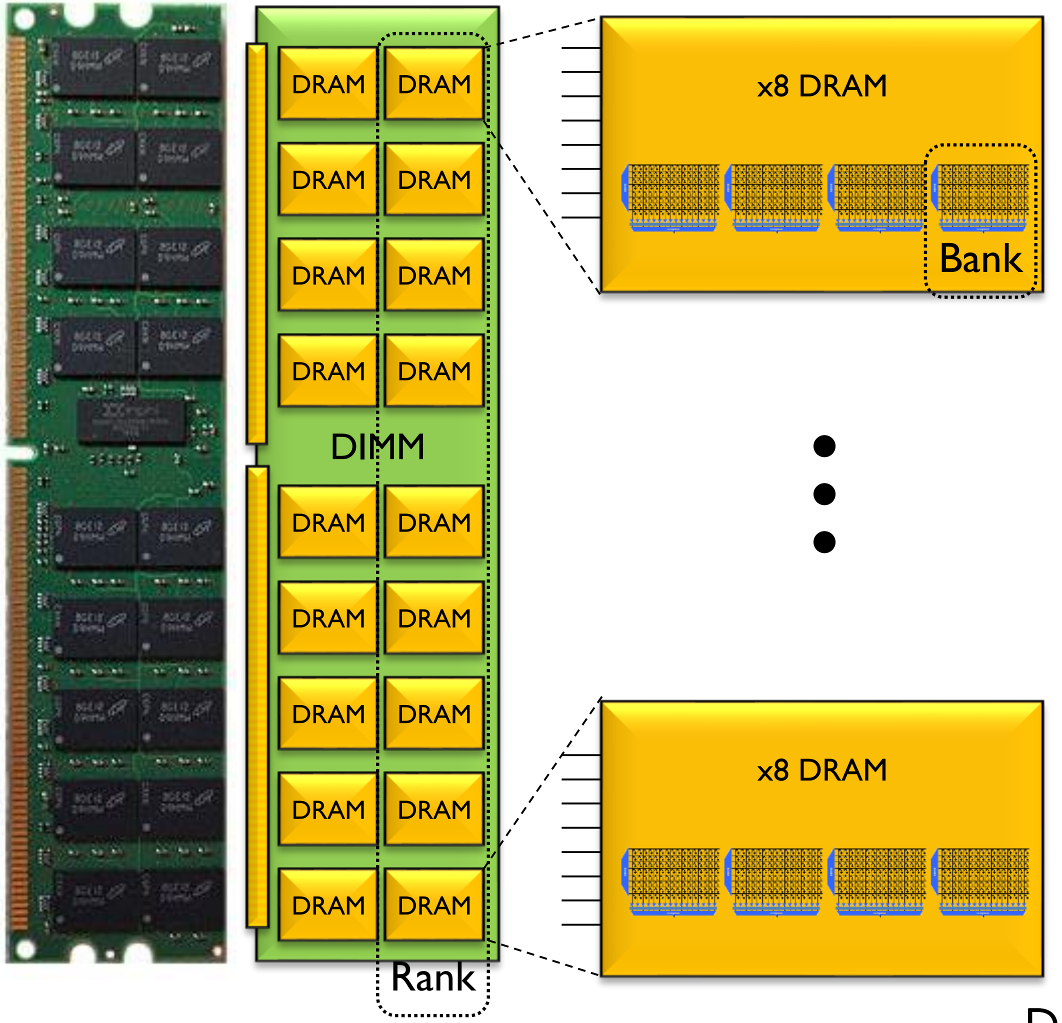
- A、存储器总线的时钟频率是1066Mhz
- B、芯片内部输入输出缓冲采用8位预取技术
- C、存储器器总线每秒传1066M次数据
- D、存储器总线带宽约为8.5GB每秒

存储系统性能优化

Overview of DRAM System



DRAM Organization



All banks within the rank share all address and control pins

All banks are independent, but can only talk to one bank at a time

x8 means each DRAM outputs 8 bits, need 8 chips for DDRx (64-bit)

Why 9 chips per rank? 64 bits data, 8 bits ECC

Dual-rank x8 (2Rx8) DIMM

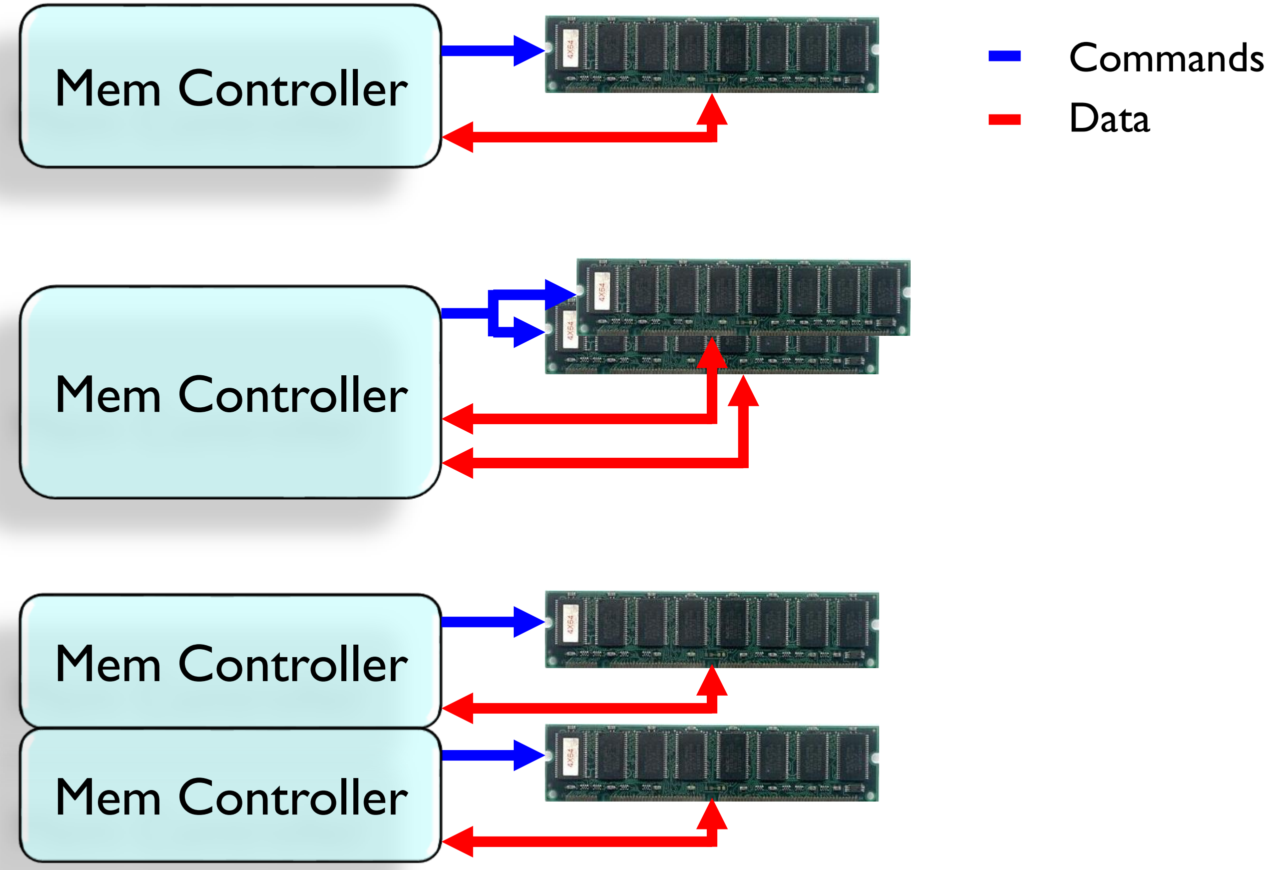
DIMM (dual in-line memory module)

Memory Channels

One controller
One 64-bit channel

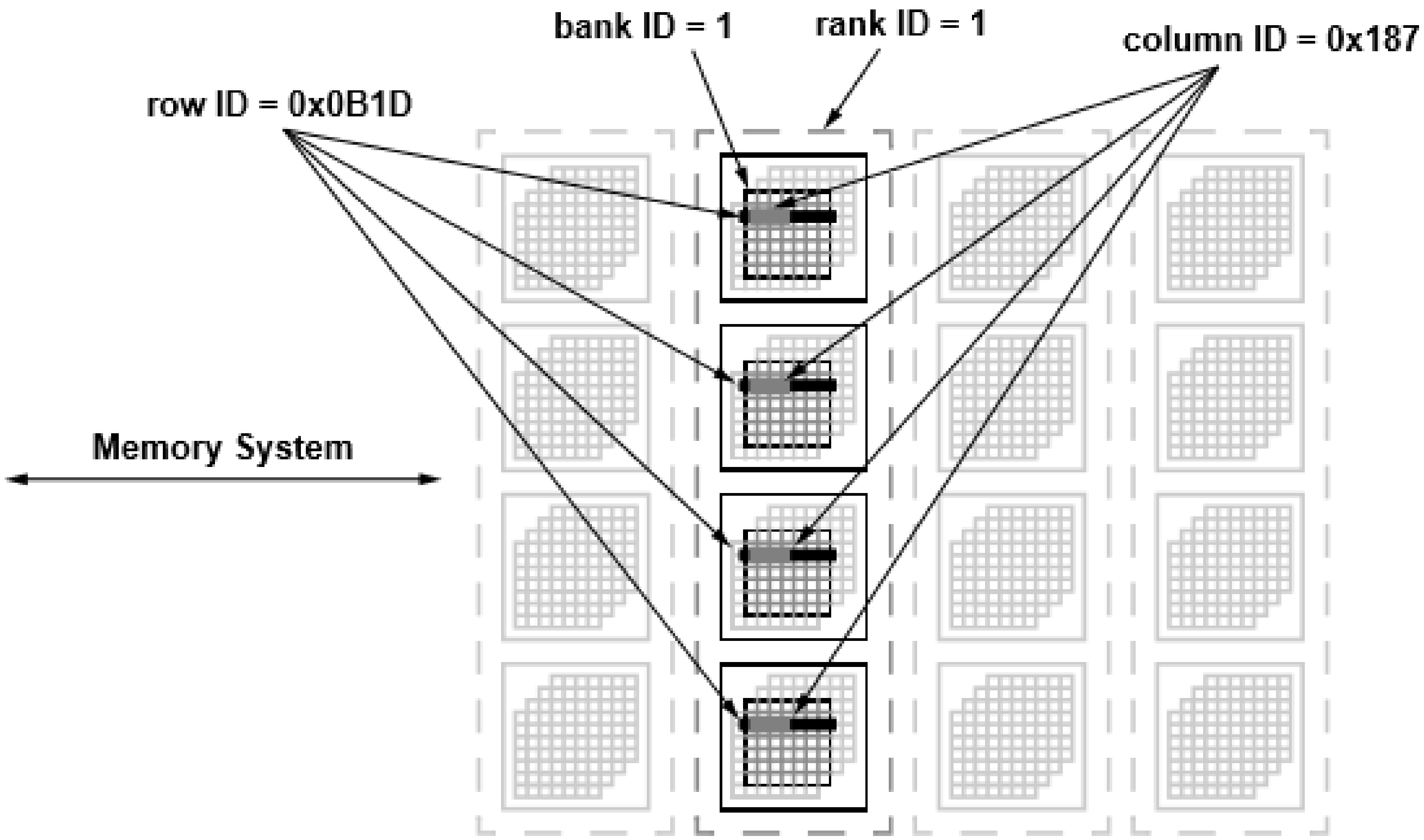
One controller
Two 64-bit channels

Two controllers
Two 64-bit channels



Use multiple channels for more bandwidth

Location of Data in a DRAM



A column of data is the smallest addressable unit of DRAM

Address Mapping

- Physical address is resolved into indices:
  - Channel ID, rank ID, bank ID, row ID, column ID
- Example address mapping policies:
  - row:rank:bank:channel:column:blkoffset
  - row:column:rank:bank:channel:blkoffset
- Consecutive cache lines can be placed in the same row to boost row buffer hit rates
- Consecutive cache lines can be placed in different ranks to boost parallelism

Address Mapping Schemes

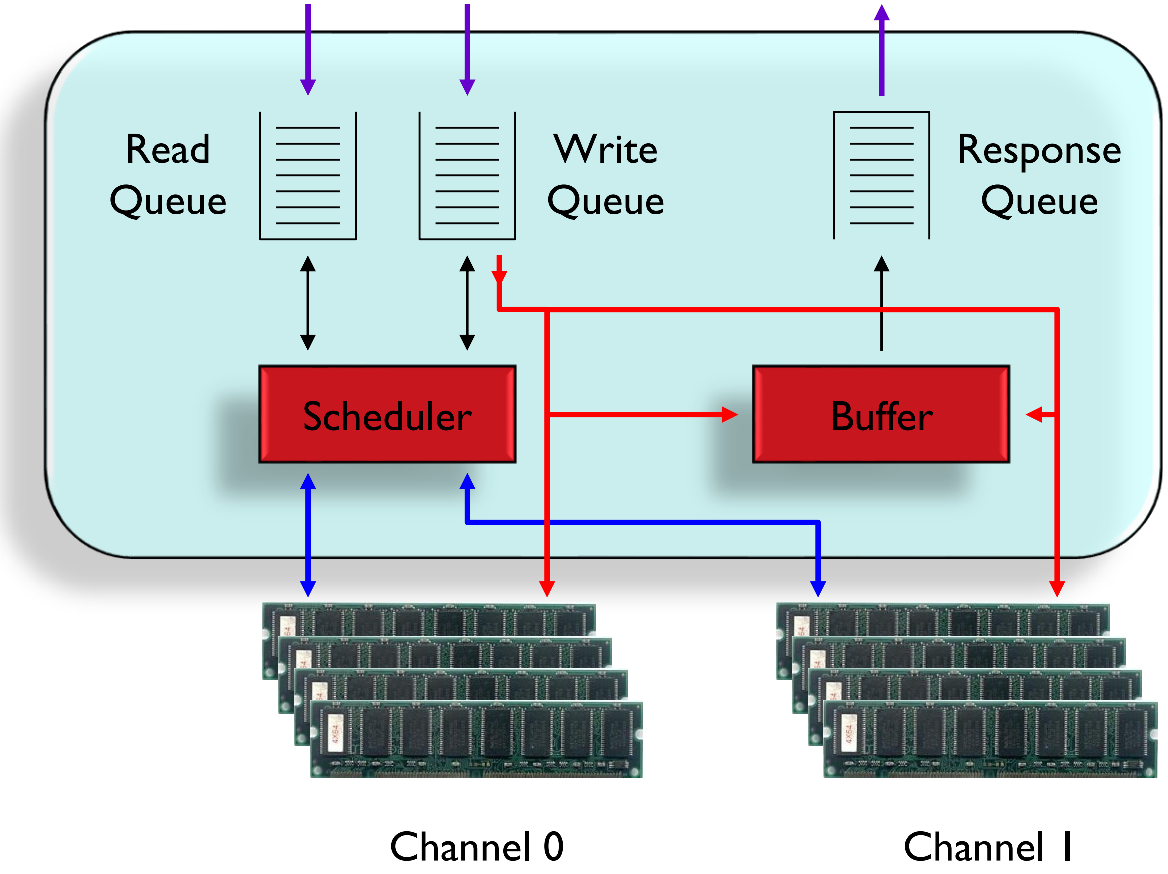
[… … … … bank column ...]

0x00000	0x00400	0x00800	0x00C00
0x00100	0x00500	0x00900	0x00D00
0x00200	0x00600	0x00A00	0x00E00
0x00300	0x00700	0x00B00	0x00F00

[… … … … column bank …]

0x00000	0x00100	0x00200	0x00300
0x00400	0x00500	0x00600	0x00700
0x00800	0x00900	0x00A00	0x00B00
0x00C00	0x00D00	0x00E00	0x00F00

Memory Controller (1/2)



Memory Controller

Memory Controller (2/2)

- Memory controller connects CPU and DRAM
- Receives requests after cache misses in LLC
  - Possibly originating from multiple cores
- Complicated piece of hardware, handles:
  - DRAM Refresh
  - Row-Buffer Management Policies
  - Address Mapping Schemes
  - Request Scheduling

Request Scheduling

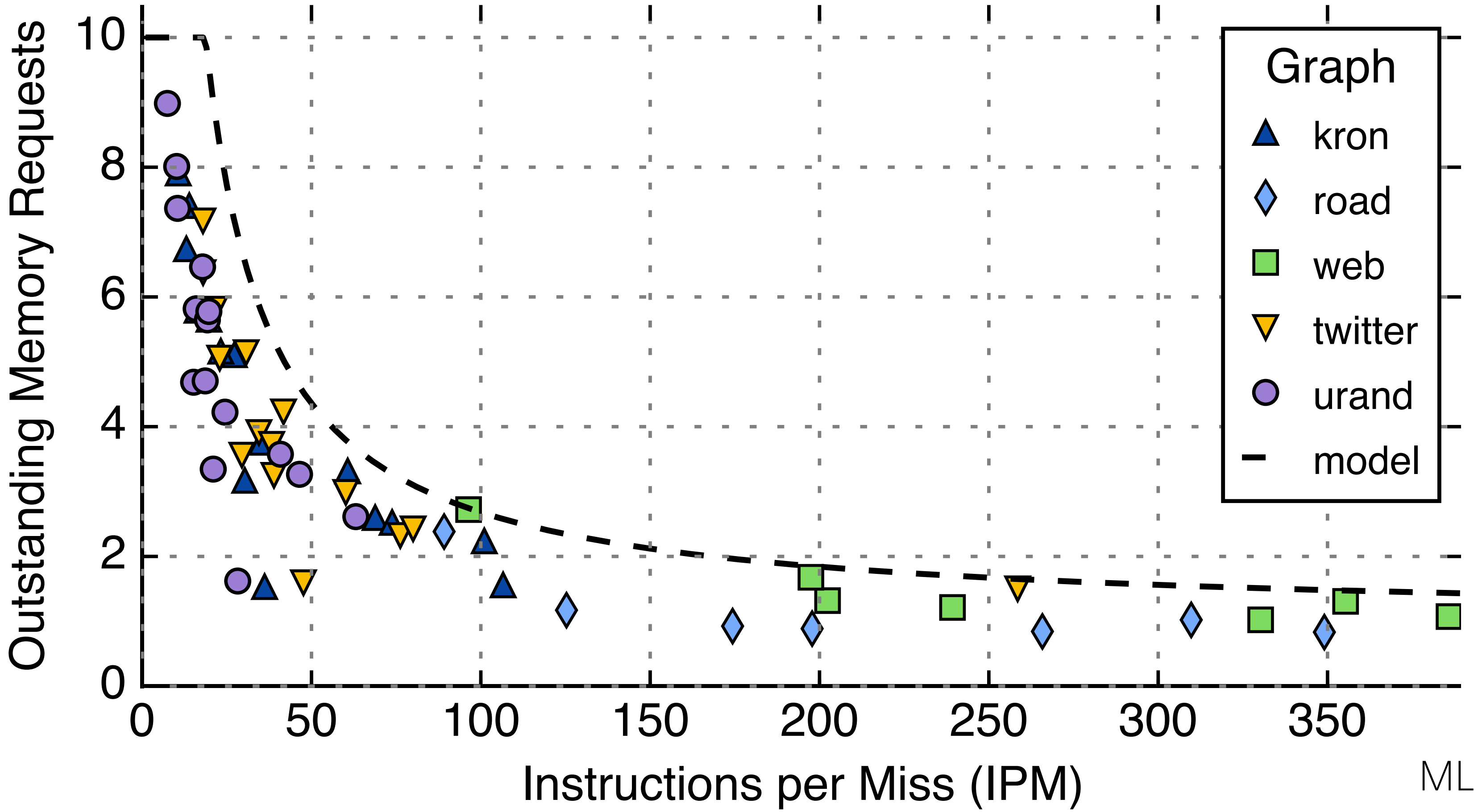
- Write buffering
  - Writes can wait until reads are done
- Queue DRAM commands
  - Usually into per-bank queues
  - Allows easily reordering ops. meant for same bank
- Common policies:
  - *First-Come-First-Served (FCFS)*
  - *First-Ready—First-Come-First-Served (FR-FCFS)*

Overcoming Memory Latency

- Caching
  - Reduce average latency by avoiding DRAM altogether
  - Limitations
    - Capacity (programs keep increasing in size)
    - Compulsory misses
- Prefetching
  - Guess what will be accessed next
    - Put in into the cache
- Memory-Level Parallelism
  - Perform multiple concurrent accesses

Discussion: MLP of Graph Workload

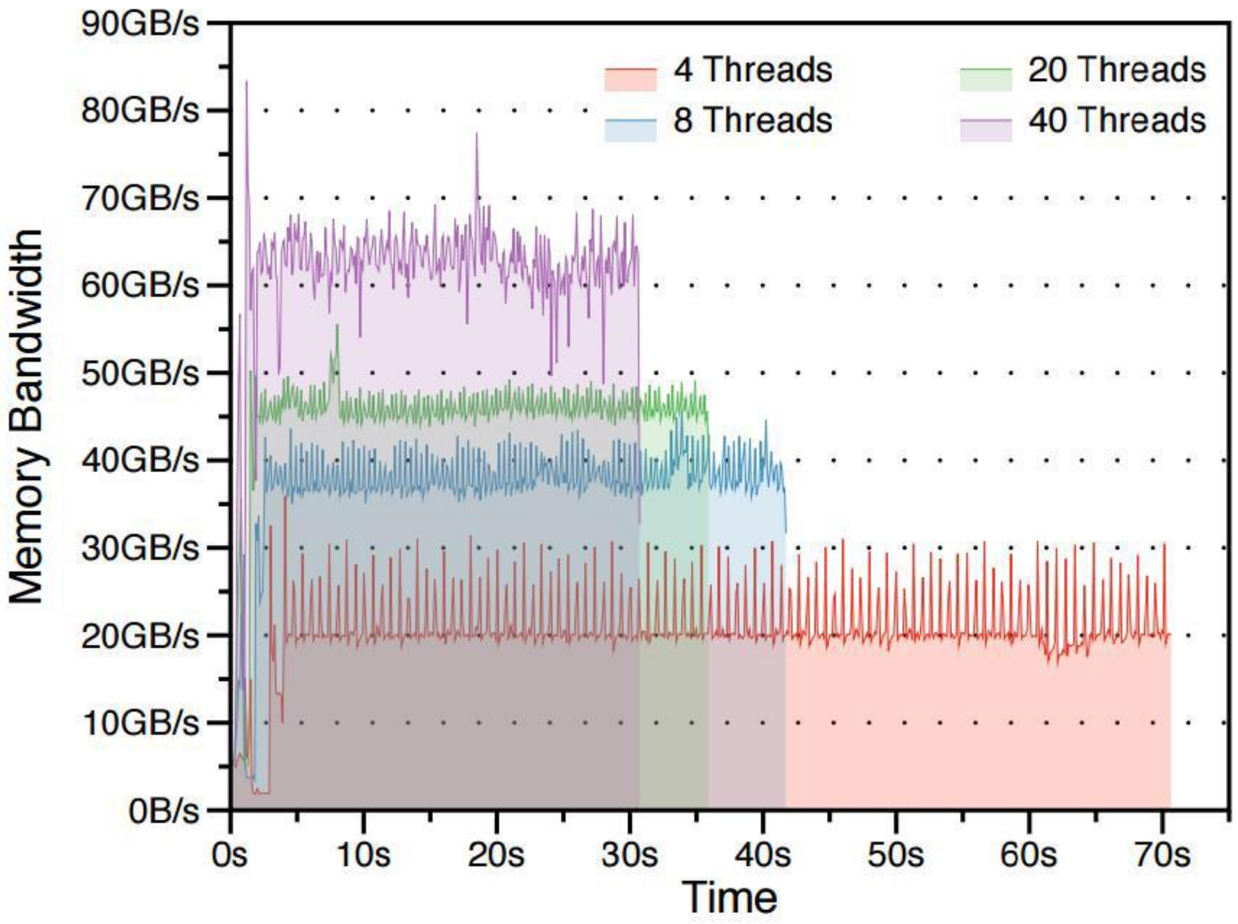
Memory bandwidth utilization limited by high IPM



MLP: (memory level parallelism)

Discussion: MLP of Graph Workload

Growing memory utilization with more CPU cores



提问时间

下一节

- 周二 16：00
- Cache
- 请做好准备

再见