


Computer Architecture
A Quantitative Approach, Sixth Edition

Chapter 2
Memory Hierarchy Design



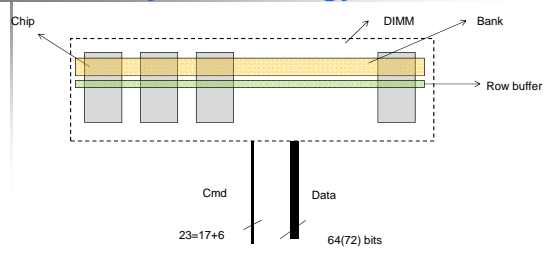
Copyright © 2019, Elsevier Inc. All rights Reserved 1

DRAM

- DIMM: Dual Inline Memory Module
- Rank: A subset of the chip on the DIMM that receive the same command (same CS)
- Chips – Banks: Banks are distributed across the chip, each request is directed to one bank.
- Array: The part of the bank in a chip is made of arrays/subarrays/mats

Copyright © 2012, Elsevier Inc. All rights reserved. 2

Memory Technology



Chip

DIMM

Bank

Row buffer

Cmd 23=17+6

Data 64(72) bits

Copyright © 2012, Elsevier Inc. All rights reserved. 3

DRAM Configuration

- Dram devices (chips) are classified by the number of data bits in each device.
- It is also classified by the data bus width (output of the chip) known as “by” × for example ×8 (read by 8) means 8 I/O pins coming out of every chip
- For example, DIMM has 64-bit data bus
 - Arranged as 8 (chips) × 8
 - Or 4 × 16
 - Or 16 × 4

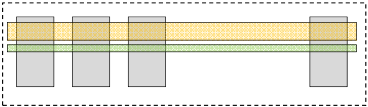
MK Copyright © 2012, Elsevier Inc. All rights reserved. 4

DRAM Operation

- Memory controller sends address and command to the DRAM (sends it to a bank)
- Memory controllers translate address to memory access address bank/row/column
- Data are sent/received on the data bus
- JEDEC protocol

MK Copyright © 2012, Elsevier Inc. All rights reserved. 5

DRAM



The diagram shows a grid of memory cells. A horizontal row of cells is highlighted in yellow, representing a row buffer. Below this row, a dashed box encloses the cells, indicating the scope of a memory access. The grid is composed of several columns and rows of cells.

- Memory access is directed to a bank
- Each access reads a row into the row buffer (say 8Kb)
 - Precharge the row (close the previous row)
 - Read data into row buffer
 - Send part of it to I/O pins

MK Copyright © 2012, Elsevier Inc. All rights reserved. 6

DRAM

- How fast
 - If the access is to an open page (data already in the row buffer) MUX to send it to the output ≈ 20 nsec
 - If the page is closed, read page into row buffer and send data to pins ≈ 40 nsec
 - If another page is open, precharge, read, send data to output $\approx 60-80$ nsec
- What to do with the page after read (page policy)?

MK Copyright © 2012, Elsevier Inc. All rights reserved. 7

Arrays

16Mb mat 4k x 4k

12-bit Row Address
RAS

12-bit Column Address
CAS

Sense amplifier

MK Copyright © 2012, Elsevier Inc. All rights reserved. 8

Page Policy

- Open Page
- Closed Page
- Some combination
- Memory Controller

MK Copyright © 2012, Elsevier Inc. All rights reserved. 9

DRAM issues

- Rearranging reads and writes
- Address mapping policy
- Scheduling Policy
- Refresh
- Error Correction

MK Copyright © 2012, Elsevier Inc. All rights reserved. 10

Avoiding memory banks Conflicts

- Suppose that we have 128 banks, and we will store 512x512 array.
- All the elements of a row will be mapped to the same bank (conflicts if we access a row).
- Usually, the number of banks is a power of 2, in this case
- Bank number = address MOD number of banks
- Address within a bank = Address/Number of banks
- This is a trivial calculation if the number of banks is a power of 2.
- If the number of memory banks is a prime number, that will decrease conflicts, but division and MOD will be very expensive

MK Copyright © 2012, Elsevier Inc. All rights reserved. 11

Avoiding memory Banks Conflicts

- MOD can be calculated very efficiently if the prime number is 1 less than a power of 2.
- Division still a problem
- But if we change the mapping such that
- Address in a bank = address MOD number of words in a bank.
- Since the number of words in a bank is usually a power of 2, that will lead to a very efficient implementation.
- Consider the following example, the first case is the usual 4 banks, then 3 banks with sequential interleaving and modulo interleaving and notice the conflict free access to rows and columns of a 4 by 4 matrix

MK Copyright © 2012, Elsevier Inc. All rights reserved. 12

Example

Add in a bank					SE	Q		M	O	D
	0	1	2	3	0	1	2	0	1	2
0	0	1	2	3	0	1	2	0	16	8
1	4	5	6	7	3	4	5	9	1	17
2	8	9	10	11	6	7	8	18	10	2
3	12	13	14	15	9	10	11	3	19	11
4	16	17	18	19	12	13	14	12	4	20
5	20	21	22	23	15	16	17	21	13	5
6	24	25	26	27	18	19	20	6	22	14
7	28	29	30	31	21	22	23	15	7	23

MK Copyright © 2012, Elsevier Inc. All rights reserved. 13

Memory Dependability

- Memory is susceptible to cosmic rays
- Soft errors*: dynamic errors
 - Detected and fixed by error correcting codes (ECC)
- Hard errors*: permanent errors
 - Use spare rows to replace defective rows
- Chipkill: a RAID-like error recovery technique

MK Copyright © 2012, Elsevier Inc. All rights reserved. 14
