



Chapter 2

Memory Hierarchy Design

DRAM

Memory Technology

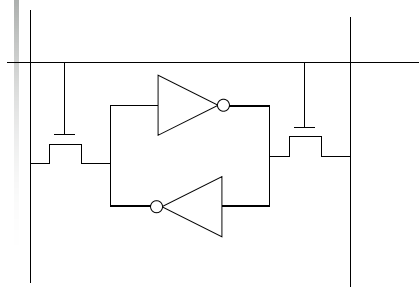
- Performance metrics
 - Latency is concern of cache
 - Bandwidth is concern of multiprocessors and I/O
 - Access time
 - Time between read request and when desired word arrives
 - Cycle time
 - Minimum time between unrelated requests to memory
- DRAM used for main memory, SRAM used for cache

Memory Technology

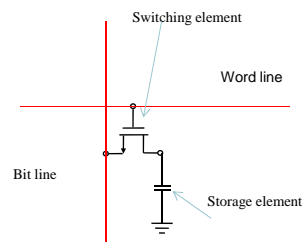
- SRAM
 - Requires low power to retain bit
 - Requires 6 transistors/bit

- DRAM
 - Must be re-written after being read
 - Must also be periodically refreshed
 - Every ~ 8 ms
 - Each row can be refreshed simultaneously
 - One transistor/bit
 - Address lines are multiplexed:
 - Upper half of address: row access strobe (RAS)
 - Lower half of address: column access strobe (CAS)

DRAM vs SRAM



SRAM

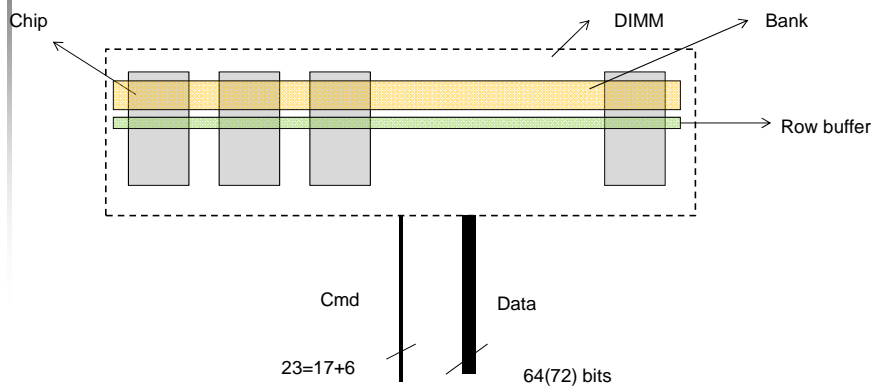


DRAM

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

Memory Technology



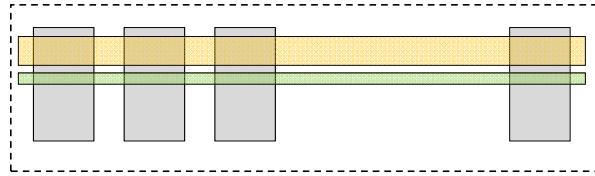
DRAM Configuration

- 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

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

DRAM



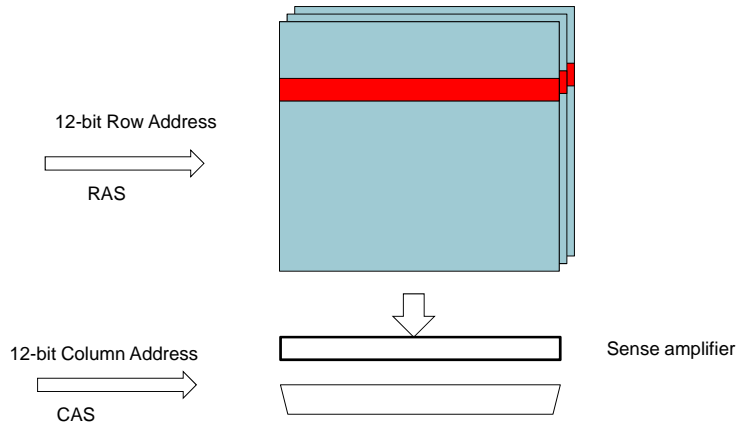
- Memory access is directed to a bank
- Each access reads a row into the row buffer (say 8Kb)
 - Precharge the row (close the previous row)
 - Read data into row buffer
 - Send part of to to I/O pins

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)?

Arrays

16Mb mat 4k x 4k



Page Policy

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

DRAM issues

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

Memory Technology

- Amdahl:
 - Memory capacity should grow linearly with processor speed
 - Unfortunately, memory capacity and speed has not kept pace with processors
- Some optimizations:
 - Multiple accesses to same row
 - Synchronous DRAM
 - Added clock to DRAM interface
 - Burst mode with critical word first
 - Wider interfaces
 - Double data rate (DDR)
 - Multiple banks on each DRAM device

Memory Optimizations

Production year	Chip size	DRAM Type	Row access strobe (RAS)		Column access strobe (CAS)/ data transfer time (ns)	Cycle time (ns)
			Slowest DRAM (ns)	Fastest DRAM (ns)		
1980	64K bit	DRAM	180	150	75	250
1983	256K bit	DRAM	150	120	50	220
1986	1M bit	DRAM	120	100	25	190
1989	4M bit	DRAM	100	80	20	165
1992	16M bit	DRAM	80	60	15	120
1996	64M bit	SDRAM	70	50	12	110
1998	128M bit	SDRAM	70	50	10	100
2000	256M bit	DDR1	65	45	7	90
2002	512M bit	DDR1	60	40	5	80
2004	1G bit	DDR2	55	35	5	70
2006	2G bit	DDR2	50	30	2.5	60
2010	4G bit	DDR3	36	28	1	37
2012	8G bit	DDR3	30	24	0.5	31

Figure 2.13 Times of fast and slow DRAMs vary with each generation. (Cycle time is defined on page 95.) Performance improvement of row access time is about 5% per year. The improvement by a factor of 2 in column access in 1986 accompanied the switch from NMOS DRAMs to CMOS DRAMs. The introduction of various burst transfer modes in the mid-1990s and SDRAMs in the late 1990s has significantly complicated the calculation of access time for blocks of data; we discuss this later in this section when we talk about SDRAM access time and power. The DDR4 designs are due for introduction in mid- to late 2012. We discuss these various forms of DRAMs in the next few pages.

Memory Optimizations

Standard	Clock rate (MHz)	M transfers per second	DRAM name	MB/sec /DIMM	DIMM name
DDR	133	266	DDR266	2128	PC2100
DDR	150	300	DDR300	2400	PC2400
DDR	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	1066–1600	2133–3200	DDR4-3200	17,056–25,600	PC25600

Figure 2.14 Clock rates, bandwidth, and names of DDR DRAMs and DIMMs in 2010. 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. Although not shown in this figure, DDRs also specify latency in clock cycles as four numbers, which are specified by the DDR standard. For example, DDR3-2000 CL 9 has latencies of 9-9-9-28. What does this mean? With a 1 ns clock (clock cycle is one-half the transfer rate), this indicate 9 ns for row to columns address (RAS time), 9 ns for column access to data (CAS time), and a minimum read time of 28 ns. Closing the row takes 9 ns for precharge but happens only when the reads from that row are finished. In burst mode, transfers occur on every clock on both edges, when the first RAS and CAS times have elapsed. Furthermore, the precharge is not needed until the entire row is read. DDR4 will be produced in 2012 and is expected to reach clock rates of 1600 MHz in 2014, when DDR5 is expected to take over. The exercises explore these details further.

Memory Optimizations

- 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

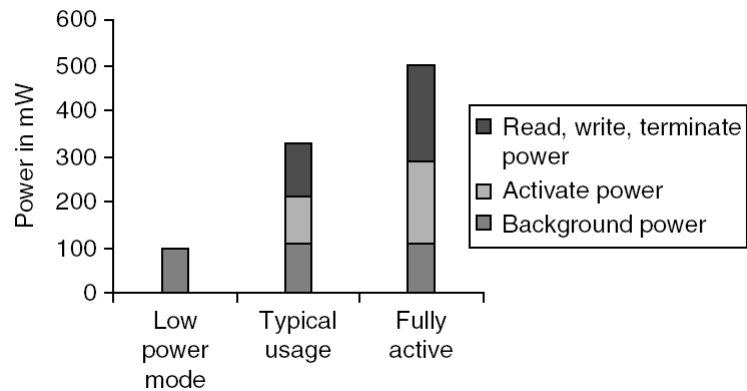
- GDDR5 is graphics memory based on DDR3

Memory Optimizations

- Graphics memory:
 - Achieve 2-5 X bandwidth per DRAM vs. DDR3
 - Wider interfaces (32 vs. 16 bit)
 - Higher clock rate
 - Possible because they are attached via soldering instead of socketed DIMM modules

- Reducing power in SDRAMs:
 - Lower voltage
 - Low power mode (ignores clock, continues to refresh)

Memory Power Consumption



Flash Memory

- Type of EEPROM
- Must be erased (in blocks) before being overwritten
- Non volatile
- Limited number of write cycles
- Cheaper than SDRAM, more expensive than disk
- Slower than SRAM, faster than disk

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

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

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

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