UT 3. Memory subsystem

Tema 3.3 Main Memory Performance Optimizations

A. Doménech, J. Duato, P. López, V. Lorente,
A. Pérez, S. Petit, J.C. Ruiz, S. Sáez, J. Sahuquillo

Department of Computer Engineering
Universitat Politècnica de València

Contents

1. Memory technology and performance model
2. Enhancing SDRAM performance

Bibliography

John L. Hennessy and David A. Patterson.
Computer Architecture, Fifth Edition: A Quantitative Approach.
Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 5 edition, 2012.

Contents

1 Memory technology and performance model

2 Enhancing SDRAM performance

Concepts

Main memory supplies the requests of the cache and I/O subsystem.

Performance goal

From the cache point of view, the goal is to reduce miss penalty (MP).

Performance metrics

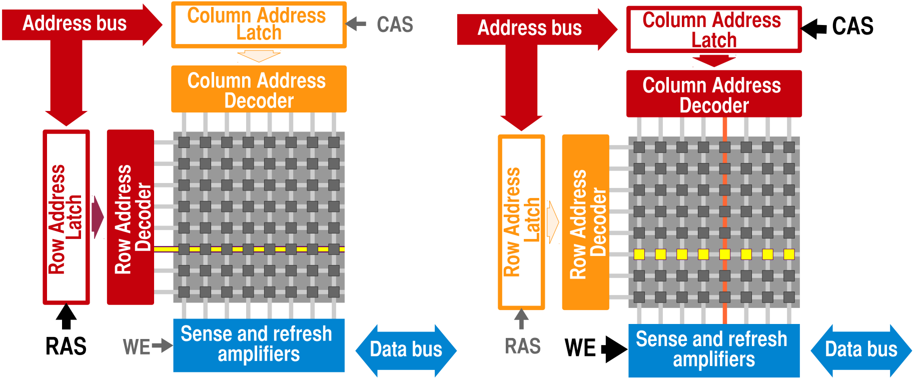
If a single data is accessed $\rightarrow MP =$ Memory access time
If multiple data are accessed (e.g. a cache block consisting of B words): $\rightarrow MP = L + \frac{1}{B_w} B$

- L, Latency: Time to satisfy the first access.
- B_w, Bandwidth: Number of words transferred per time unit.

It is easier to increase the bandwidth than to reduce the latency because, for a given VLSI technology, increasing memory size leads to increasing latency.

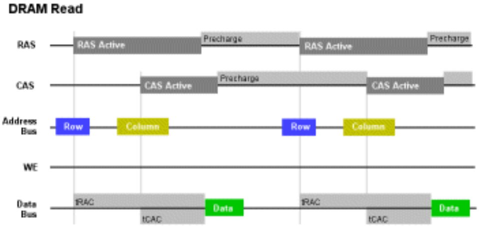
Evolution of DRAM technology

- *Traditional DRAM*. Due to pin count constraints, the address is multiplexed. First the row address is transmitted (RAS signal is activated to validate the address), then the column address (CAS signal is activated to validate it).



Evolution of DRAM technology (cont.)

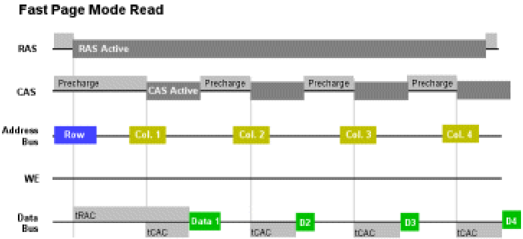
Timing:



- A single memory word is read/written every time memory is accessed. However, an entire row is internally read every time a word is accessed, later refreshing it.
- After accessing a word, next access cannot start until the memory cycle is complete → precharge refreshes and closes a row.

Evolution of DRAM technology (cont.)

- *Fast page mode*. If a row buffer is added, accesses to other words in the same row will be faster.
- Once the row is available, several column addresses can be read (or written) in sequence.
- Only the column address is required for each access.



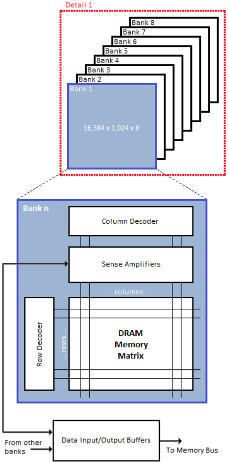
Current DRAM technology

Synchronous DRAM (SDRAM) characteristics:

- SDRAMs are synchronous:
  - The clock signal is sent to memory
  - The clock frequency is defined by the memory controller
  - Timing is measured in clock cycles. The number of cycles required for each operation (sending addresses, accessing and transferring data) are read from a ROM in the SDRAM module and are used to configure the memory controller.

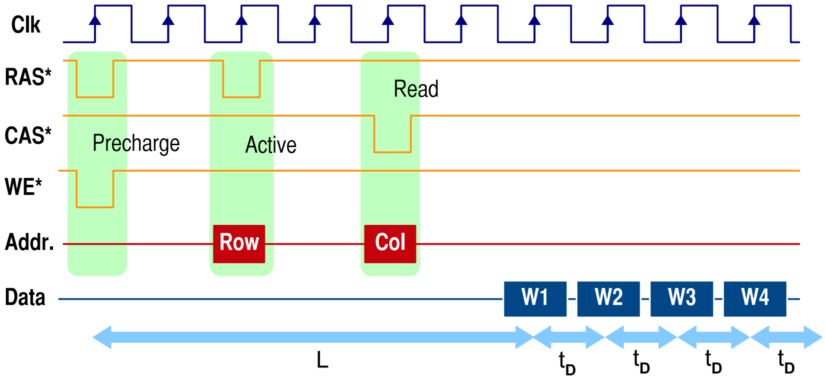
Current DRAM technology (cont.)

- SDRAM chips are organized as one or more banks
  - Each bank is a memory cell array.
  - Once a bank row is activated, it is possible to read and/or write any column from it.
- Burst mode: SDRAMs use a self incrementing counter and a mode register to determine the column address sequence after the first access to a row. This allows faster DRAM operations since the time to set up subsequent column addresses is removed.



Current DRAM technology (cont).

SDRAM read chronogram



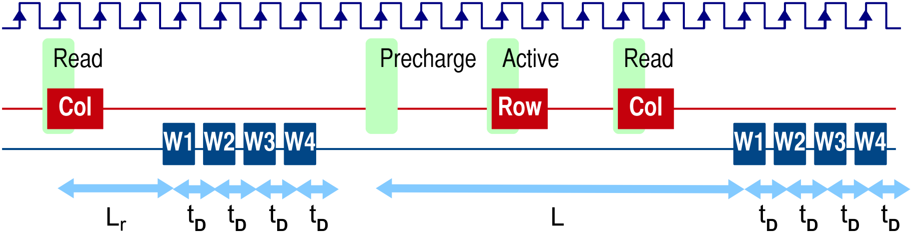
Current DRAM technology (cont).

Implications:

- Cache and disk blocks are accessed in burst mode so that, once the first word is accessed, the remaining ones could be transferred at a high rate.
- A higher clock frequency increases data transfer rate but it does not reduce memory access time *L* for the first word in a burst.
- Memory access time *L* depends on memory timing parameters, which are specified as an integer number of clock cycles: the lowest number of cycles that allow the operation to complete. Thus, due to rounding effects, it may happen that a higher clock frequency leads to longer access time.
- As many rows as chip banks can be active at a time.
- Miss penalty depends on whether consecutive block accesses belong to the same row. This introduces variability in the miss penalty, making it dependent on memory access patterns.

Current DRAM technology (cont).

The row is already open: The row needs to be opened:



Simple memory model

Generic memory parameters:

- L: Latency or access time (time to read the first word).
- t_D: Time to transfer each word.
- B_w: Bus bandwidth, measured in words/sec.

Consider a block size of B memory words. The miss penalty MP is:

$$MP \text{ (in seconds)} = L + t_D \cdot B = L + \frac{1}{B_w} \cdot B$$

Simple memory model (cont.)

MP can also be expressed in clock cycles:

- f: Bus clock frequency.
- L_c: Latency, measured in cycles at frequency f.
- B_{wc}: Bus bandwidth, measured in words/cycle at frequency f.

$$MP \text{ (in cycles)} = L_c + \frac{1}{B_{wc}} \cdot B$$

$$MP \text{ (in seconds)} = MP \text{ (in cycles)} / f$$

Simple memory model (cont.)

General case (the requested row may not be open):

- Let *ML (memory locality)* be the probability that a cache miss requests a memory block that belongs to one of the open (activated) rows.
- In that case, access time is shortened since the corresponding row is already stored in a row buffer.
- Let L_r be the reduced access time or latency. Let L_{rc} be the value of L_r when measured in cycles at frequency f.

The average miss penalty is:

$$MP \text{ (in seconds)} = L \cdot (1 - ML) + L_r \cdot ML + \frac{1}{B_w} \cdot B = MP \text{ (in cycles)}/f$$

$$MP \text{ (in cycles)} = L_c \cdot (1 - ML) + L_{rc} \cdot ML + \frac{1}{B_{wc}} \cdot B$$

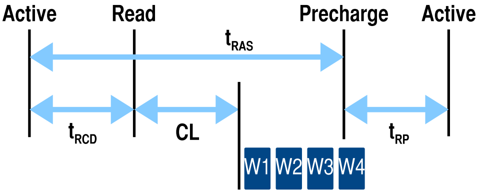
Computation of L and L_r: Memory timing parameters

Timing in commercial SDRAMs is defined by the clock frequency and by four timing parameters, separated by dashes. Those parameters, in order of appearance, are the following:

- CL: The time (in cycles) between sending a column address to the memory and the beginning of the data burst.
  This is the time it takes to read the first bit of memory from a DRAM with the correct row already open.
- t_{RCD}: The number of clock cycles required between the opening of a row of memory and accessing columns within it.
  The time to read the first bit of memory from a DRAM without an active row is $t_{RCD} + CL$.

Computation of L and L_r: Memory timing parameters (cont.)

- t_{RP}: The number of clock cycles required between the issuing of the precharge command and opening the next row.
  The time to read the first bit of memory from a DRAM with the wrong row open is $t_{RP} + t_{RCD} + CL$.
- t_{RAS}: The number of clock cycles required between a bank active command and issuing the precharge command.
  This is the time needed to internally refresh the row, and overlapping with t_{RCD}. This fourth parameter is sometimes dropped when specifying memory timings.

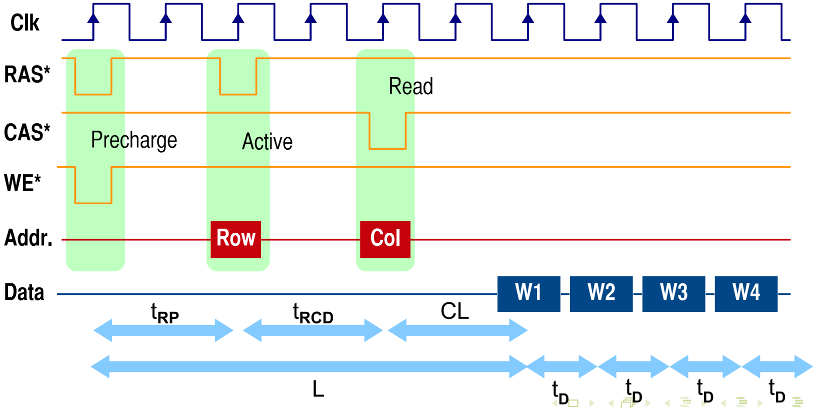


Computation of L and L_r: Memory timing parameters (cont.)

Computation of L and L_r from the timing parameters:

$$L_c = t_{RP} + t_{RCD} + CL \text{ cycles}$$

$$L_{rc} = CL \text{ cycles}$$



Example

HyperX KHX1600C9D3/4G is a 512M x 64-bit (4GB) DDR3-1600 (800 MHz bus clock, transferring two words per clock cycle) SDRAM. It can run at a low latency timing of 9-9-9-27 at 1.65V.

Compute the *MP* for a block size of $B = 8$, assuming $ML = 0$.

$$L_c = t_{RP} + t_{RCD} + CL = 9 + 9 + 9 = 27 \text{ cycles}$$

Since it is DDR (see slide 27), the bus bandwidth is $B_{wc} = 2$ words/cycle. Thus, for $ML = 0$:

$$MP \text{ (in cycles)} = L_c + \frac{1}{B_{wc}} \cdot B = 27 + 8/2 = 31 \text{ cycles}$$

$$MP \text{ (in seconds)} = MP \text{ (in cycles)}/f = 31/0.8 \text{ GHz} = 38.75 \text{ ns}$$

Example (cont.)

If it was a SDR SDRAM ($B_{wc} = 1$), we would have:

$$MP \text{ (in cycles)} = L_c + \frac{1}{B_{wc}} \cdot B = 27 + 8/1 = 35 \text{ cycles}$$

$$MP \text{ (in seconds)} = MP \text{ (in cycles)}/f = 35/0.8 \text{ GHz} = 43.75 \text{ ns}$$

Contents

1 Memory technology and performance model

2 Enhancing SDRAM performance

Techniques to enhance SDRAM performance

$$MP = L \cdot (1 - ML) + L_r \cdot ML + \frac{1}{B_w} \cdot B$$

MP can be reduced by reducing L and B, and by increasing B_w and ML. However:

- L is by far the largest contributor to MP.
- L remains roughly constant when increasing $f \rightarrow$ except for rounding effects, L_c increases linearly with f.
- ML mostly depends on memory access patterns, but it also depends on the number of banks.

Techniques to enhance SDRAM performance (cont.)

Techniques to improve the performance of main memory:

- Increase the bus width: $B \downarrow$.
- Increase B_{wc} while keeping f constant by transferring data at both the rising and falling edges of the clock signal: $B_w \uparrow$.
- Increase the clock frequency f while keeping B_{wc} constant:$B_w \uparrow$.
- Increase the number of memory banks: $ML \uparrow$.
- Implement several memory controllers. Although this does not reduce miss penalty (unless addresses are interleaved), it allows several misses to be concurrently serviced. It also increases the total number of memory banks: $ML \uparrow$.

Increasing the memory bus width

By making the memory bus wider, more than one word can be transferred at the same time → The number of transfers is reduced.

Example: MP with a bus twice as wide in the DDR3-1600

Half the number of transfers is required. The new block size, referred to the wider bus, is $B' = \frac{B}{2} = \frac{8}{2} = 4$.

$$MP \text{ (in cycles)} = L_c + \frac{1}{B_{wc}} \cdot B' = 27 + 4/2 = 29 \text{ cycles}$$

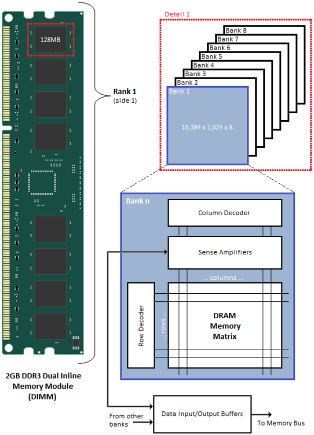
Constraints

Since L_c is by far the largest contributor to MP, little savings can be achieved by making the bus wider (for the original bus width, MP was 31 cycles, as seen in slide 20).

2. Enhancing SDRAM performance

Increasing the memory bus width (cont.)

- Mainly due to manufacturing reasons (number of pins), current memory buses are 64-bit (8-byte) wide.
- Memories are arranged as modules *Dual Inline Memory Module* (DIMMs), each of them having one or more *ranks*.
- A rank consists of enough chips to complete 64 bits.

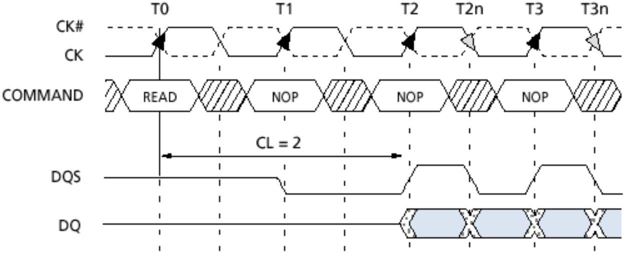


DDR: Double Data Rate

- Simple idea: Transmit data at both the rising and falling edge of the clock signal.
- The bus works at the same speed, but the bandwidth is doubled.
- The maximum signalling frequency with respect to SDR *(Single Data Rate)* remains unchanged, thus being implementable without having to enhance technology.
- Internally, the number of accessed columns is doubled (*2n-prefetch*) as well as the width of the bus connecting memory banks with the data bus → the internal clock frquency does not change.

→ B_{wc} is doubled.

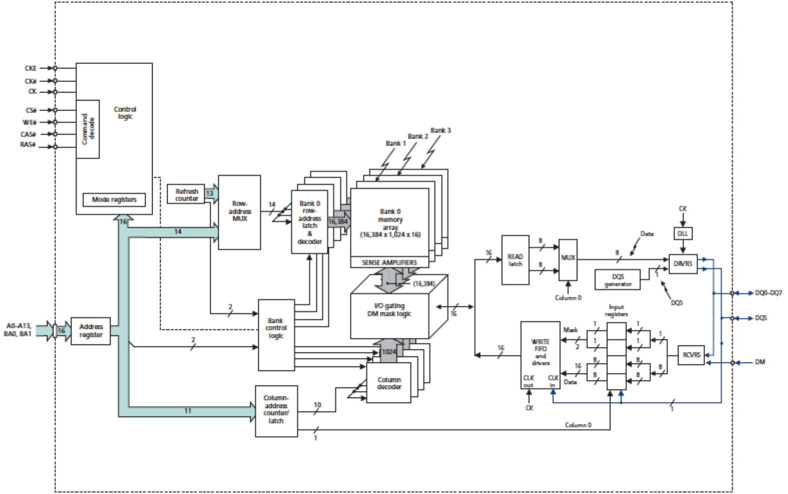
DDR: Double Data Rate (cont.)



2. Enhancing SDRAM performance

DDR 1Gbit

Figure 4: 128 Meg x 8 Functional Block Diagram



Increasing the bus clock frequency

- Memory bus clock frequency has been increased over time, also reducing voltage to reduce power consumption.
- Several techniques have been developed to keep signal integrity at higher clock frequencies (differential transmission, terminating resistors, replacement of buses with point-to-point links . . .).
- Clock frequency and voltage have been standardized by JEDEC. Standard values are as follows:
    - DDR.
        - 2.5V for bus clock up to 166 MHz and 2.6V for bus clock at 200 MHz.
        - Peak transfer rates up to 3200 MB/s.
        - Up to 1 Gb per chip.

Increasing the bus clock frequency (cont.)

- DDR2. The number of accessed columns is doubled with respect to DDR (*4n-prefetch*) as well as the width of the bus connecting memory banks with the data bus.
  - Reduces voltage to 1.8V.
  - Increases bus clock frequency up to 533 MHz.
  - Peak transfer rates up to 8533 MB/s.
  - Up to 4 Gb per chip.
- DDR3. The number of accessed columns is doubled with respect to DDR2 (*8n-prefetch*) as well as the width of the bus connecting memory banks with the data bus.
  - Reduces voltage to 1.5V.
  - Increases bus clock frequency up to 1066 MHz.
  - Peak transfer rates up to 17066 MB/s
  - Up to 16 Gb per chip.

Increasing the bus clock frequency (cont.)

- DDR4. Keeps *8n-prefetch*. Banks are arranged into groups that are also addressable. Each group can be independently and concurrently accessed, regardless of the state of other groups. Memory buses have been replaced with channels with point-to-point links.
  - Reduces voltage to 1.2V.
  - Increases bus clock frequency up to 1200 MHz.
  - Peak transfer rates up to 19200 MB/s.
  - Up to 16 Gb per chip (more in the future).

Increasing the bus clock frequency (cont.)

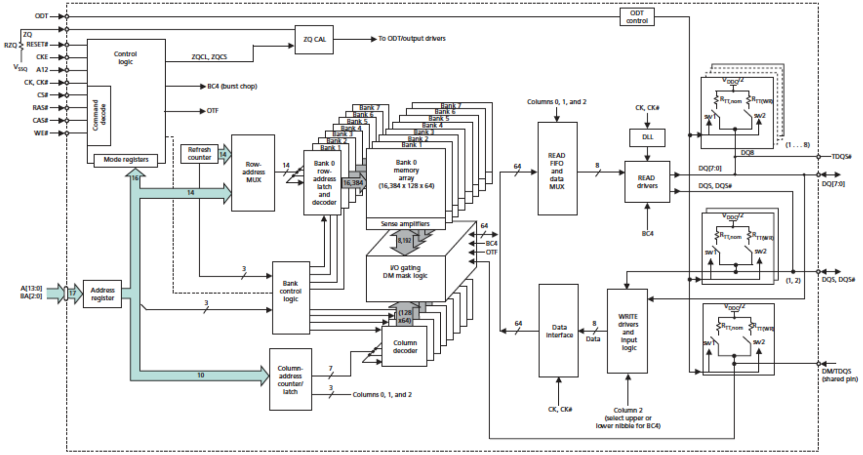
To report performance, notations are:

- DDRn-*xxxx*, where *xxxx* indicates the transfer rate in Mtransfers/s.
  - Bus clock frequency: $f = xxxx/2$ MHz.
  - Examples:
    - DDR-400 works at 200MHz and provides 400 Mtransfers/s = 3200 MB/s.
    - DDR3-1600 works at 800MHz and provides 1600 Mtransfers/s = 12800 MB/s.
- PCn-*yyyy*, where *yyyy* is the bus bandwidth in MB/s.
  - Bus clock frequency: $f = yyyy/(8 \times 2)$ MHz.
  - Examples:
    - PC-3200 delivers 3200 MB/s (200 MHz x 8 bytes x 2 (DDR)).
    - PC3-12800 delivers 12800 MB/s (800 MHz x 8 bytes x 2 (DDR)).

2. Enhancing SDRAM performance

DDR3 1Gbit:

Figure 4: 128 Meg x 8 Functional Block Diagram



2. Enhancing SDRAM performance

Increasing the bus clock frequency

Example: Fastest non-optional JEDEC standard latencies

Standard name	Bus clock (MHz)	Timings CL-t_{RCD}-t_{RP} (cycles)	L_c (cycles)	L (ns)	MP (ns) ($B = 8$)
DDR-400A	200	2.5-3-3	8.5	42.5	62.5
DDR2-800C	400	4-4-4	12	30	40
DDR2-1066E	533.33	6-6-6	18	33.75	41.25
DDR3-800D	400	5-5-5	15	37.50	47.50
DDR3-1066E	533.33	6-6-6	18	33.75	41.25
DDR3-1600H	800	9-9-9	27	33.75	38.75
DDR3-2133L	1066.67	12-12-12	36	33.75	37.50
DDR4-1600K	800	11-11-11	33	41.25	46.25
DDR4-2133P	1066.67	15-15-15	45	42.19	45.94
DDR4-2400R	1200	16-16-16	48	40	43.33

Increasing the bus clock frequency (cont.)

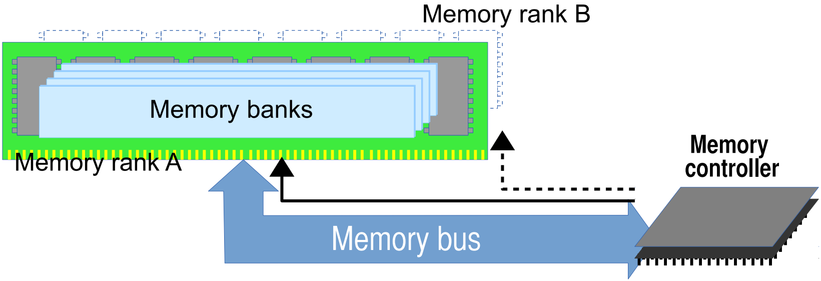
- Latency (*L*) decreased from DDR to DDR2, but it slightly increased again from DDR2 to DDR3, and from DDR3 to DDR4.
- In fact, from DDR2, latency *L* and miss penalty *MP* almost always increased with every new generation of memories (for example, DDR2-800C vs. DDR3-800D; DDR3-2133L vs. DDR4-2400R).
- But memory capacity has increased. Also, support for a higher number of cores has been added, and power consumption has been reduced.

Increasing the number of memory banks

Current SDRAM chips implement a large number of banks (typically eight or sixteen). The reasons are:

- Several rows (one per bank) can be open at a time, thus increasing *ML*.
  - To access a different open row, the bank address is supplied together with the column address.
  - Accessing a different open row is as fast as accessing the same row again.
- For a given memory capacity, increasing the number of banks reduces the bank size, thus reducing latency.
- Smaller banks also imply faster address decoders.
- A bank design can be replicated, thus simplifying SDRAM chip design.

Increasing the number of memory banks (cont.)

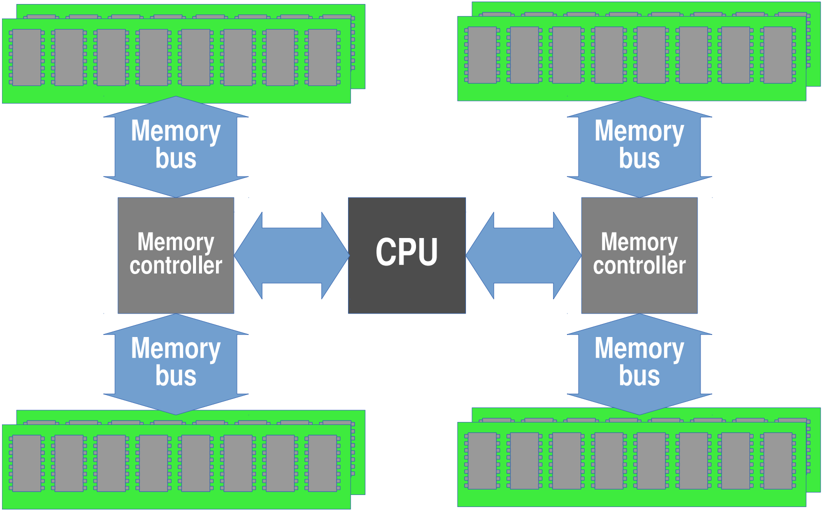


2. Enhancing SDRAM performance

Increasing the number of memory controllers

- Current high-performance processors implement multiple memory controllers.
- Each memory controller implements one or two channels to access one or more ranks of SDRAM DIMMs.
- Each DIMM implements multiple internal banks, as mentioned in the previous slide.
- The total number of rows that can be simultaneously open is the number of memory controllers times the number of channels per controller times the number of ranks per channel times the number of banks per rank.
- A large number of open rows is necessary to minimize conflicts when multiple cores concurrently initiate memory accesses. This way, many memory accesses will hit on an open row, thus maximizing *ML*.

Increasing the number of memory controllers (cont.)


Memory bus
Memory bus
Memory controller
CPU
Memory controller
Memory bus
Memory bus