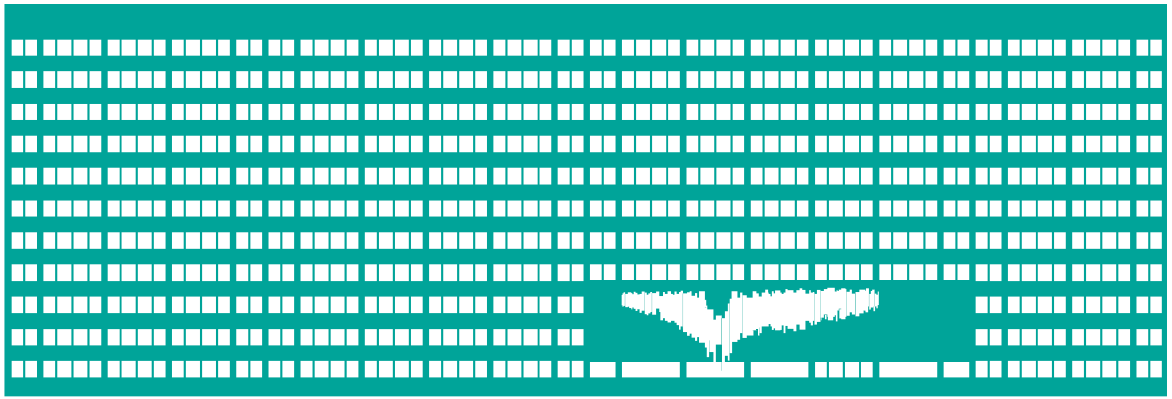
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

APPS
Architektury počítačů a paralelních systémů / Architecture of Computers and Parallel Systems

Part 05: Internal Computers Memory

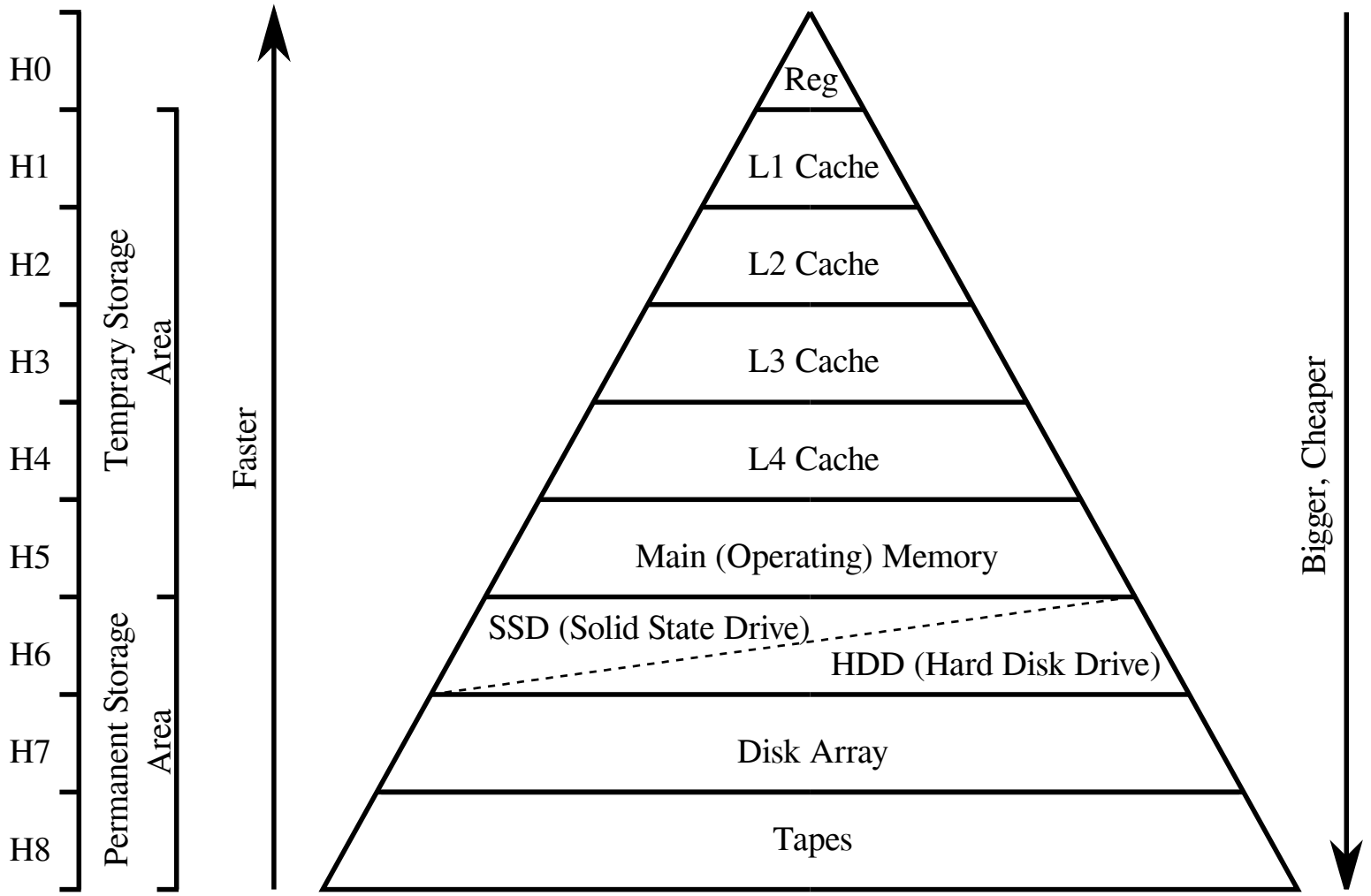
Ing. Petr Olivka, Ph.D.
Department of Computer Science,
FEECS, VSB-TUO
petr.olivka@vsb.cz
http://poli.cs.vsb.cz

History of Memories

The history of the electrical memory started with computer development after the World War II. Briefly we can summarize it in a few points:

- 1945 – Electromechanical memory realized by relay or jumpers.
- 50's – Flip-flop from tubes.
- 50's – Ferrite memory uses magnetic properties of cells.
- 50's – Magnetic drums.
- 60's – Magnetic bubble memory.
- 60's – Semiconductor memory MOS.
- 1970 – DRAM memory.
- 1971 – SRAM memory.

Memory Hierarchy



Memory Hierarchy

There are many types of memories used in today's computers. Memories are made using different technologies. There is a memory hierarchy according to their speed, capacity and price:

Registers – smallest and fastest SRAM memory in processor.

Cache L1 – SRAM memory ranging from kB to tens of kB.

Cache L2 – SRAM mem. ranging from tens of kB to tens of MB.

Main memory – DRAM size up to tens GB.

The internal memories end at this point and external memories begin.

Hard disks – magnetic memory, size up to a few TB.

Optical memory – CD, DVD, etc.

Magnetic tapes with capacity up to a few TB, but it is very slow.

Registers are the fastest and the most expensive. Tapes are the slowest and the cheapest.

Memory Classification

The internal computer memories are characterized by several parameters.

Their first division is possible by memory access:

- SAM – Serial Access Memory.
- RAM – Random Access Memory.
- Special access – stack, queue, multi-ports, associative memory.

The second division of memories can be made according to the ability to read and/or to write:

- RWM – Read and Write memory.
- ROM – Read Only Memory.
- WOM – Write Only Memory.
- Combined Memory.

… Memory Classification

The third option is to divide memories by the type of memory cell:

- DRAM – Dynamic RAM, cell is capacitor.
- SRAM – Static RAM, cell is transistor flip-flop.
- EPROM, EEPROM, Flash – programmable memory, cell is special MOS transistor.

Memories are classified by all parameters simultaneously. For example main operating memory of personal computer is usually (incorrectly) known as “RAM”. Correctly it should be marked as “RWM DRAM”.

Dynamic RAM

The Dynamic RAM has all cells realized by a tiny capacitor with one transistor. The scheme is on the next slide.

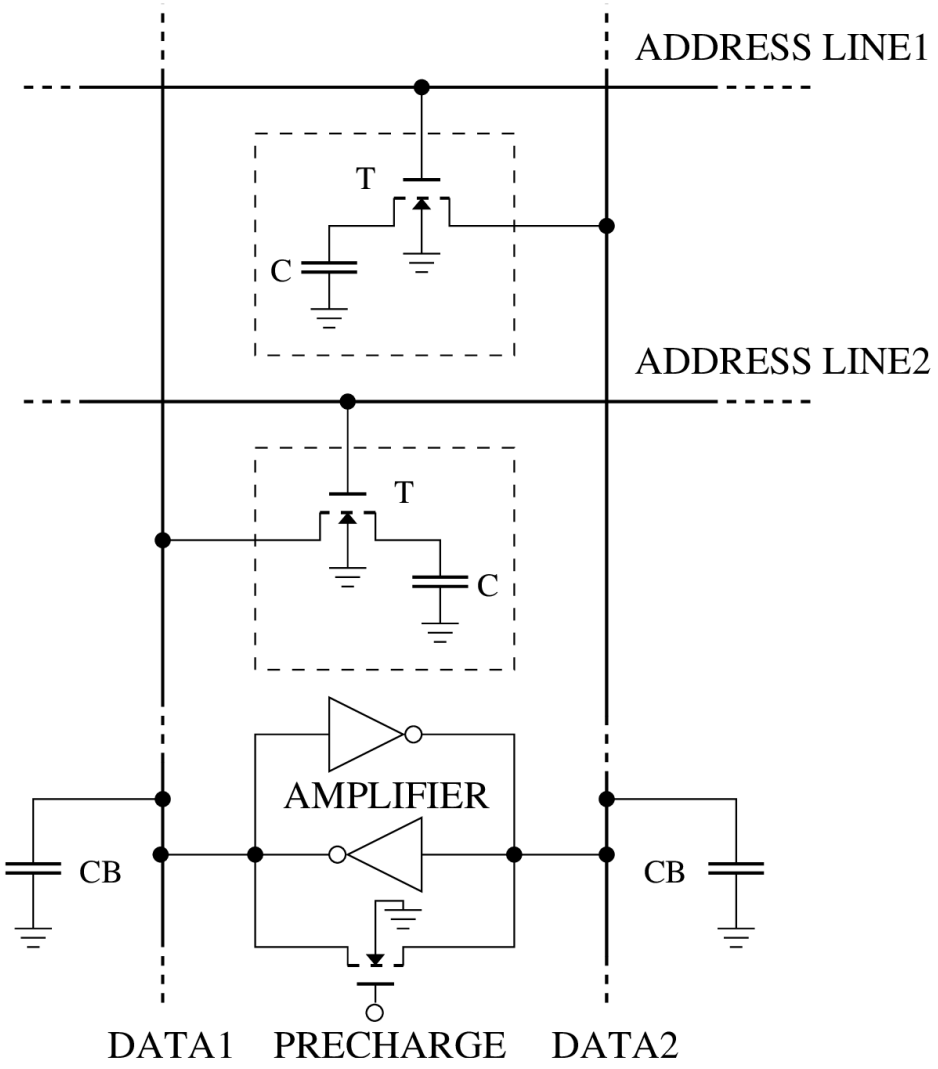
All capacitors have capacity in fF (femto Farad) and they are not able to store charge for a long time. They are very quickly losing their charge. Therefore the charge of all capacitors in the chip has to be periodically refreshed every few milliseconds. The refresh was earlier realized by reading the memory, e.g. by the DMA usage. Now the “hidden” refresh is implemented directly in the chip and does not need external circuits.

The transistor in all cells only passes charge from/to the capacitor. Every reading of cell discharges the capacitor and therefore it has to be charged back.

What follows is the circuit diagram of the DRAM memory cell implementation:

... Dynamic RAM

DRAM cell scheme.



... Dynamic RAM

The memory chip does not contain only one or a few memory cells. The chip integrates many millions or billions of cells. Cells are organized in square matrix and all cells have their own addresses, given by address of a row and address of a column. Therefore we have to specify two addresses to select a single cell: the row and the column.

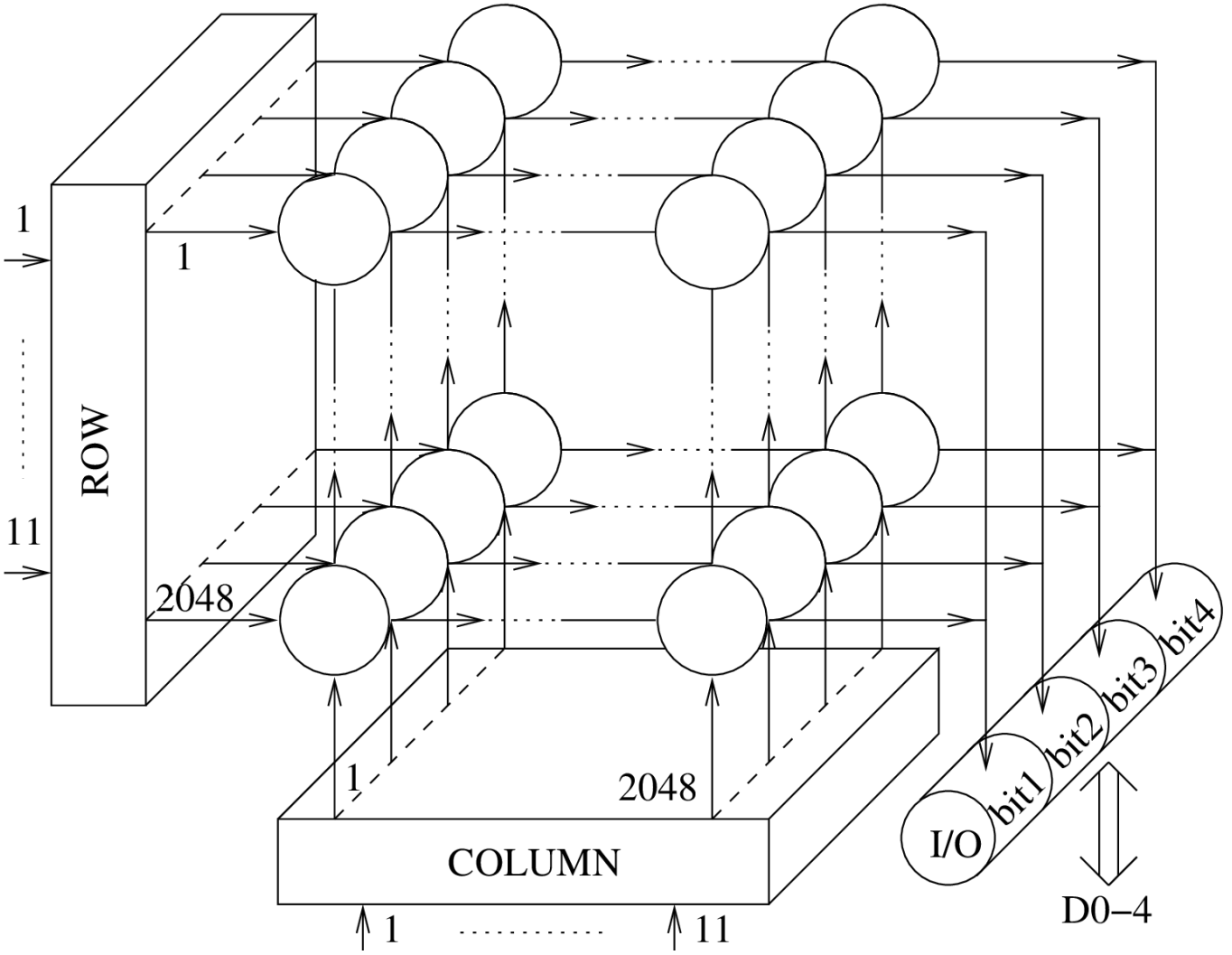
This two step addressing is a little slower than the direct addressing, but it needs less address signals. For example 1M of bits can be organized in a matrix of 1024x1024 (1k x 1k) and thus only 10 address signals are needed. Normally for 1M of bits, 20 bits would be necessary.

One matrix forms one layer on the chip. The chip can integrate more layers and on the same address of the row and the column it has more bits stored.

The scheme of 3D cells organization is on the next scheme. The chip contains 4M x 4 bits:

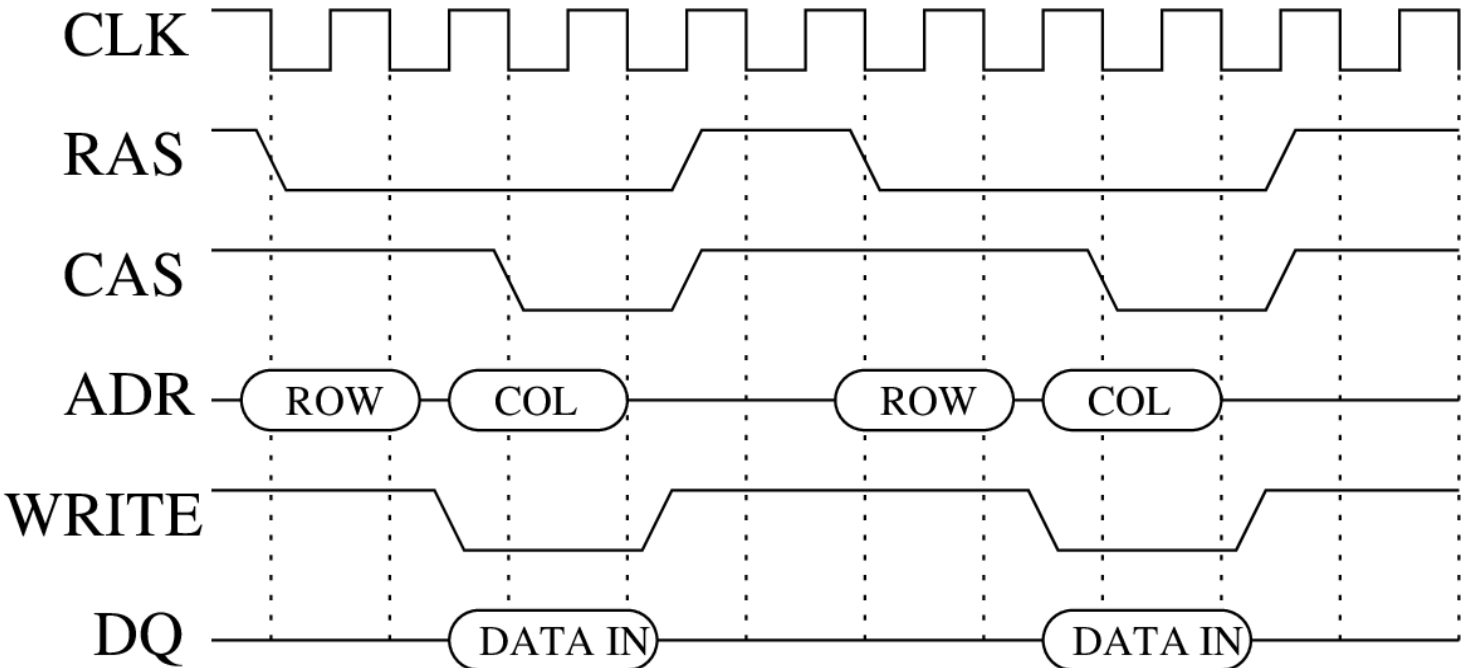
... Dynamic RAM

3D DRAM chip organization.



Write to DRAM

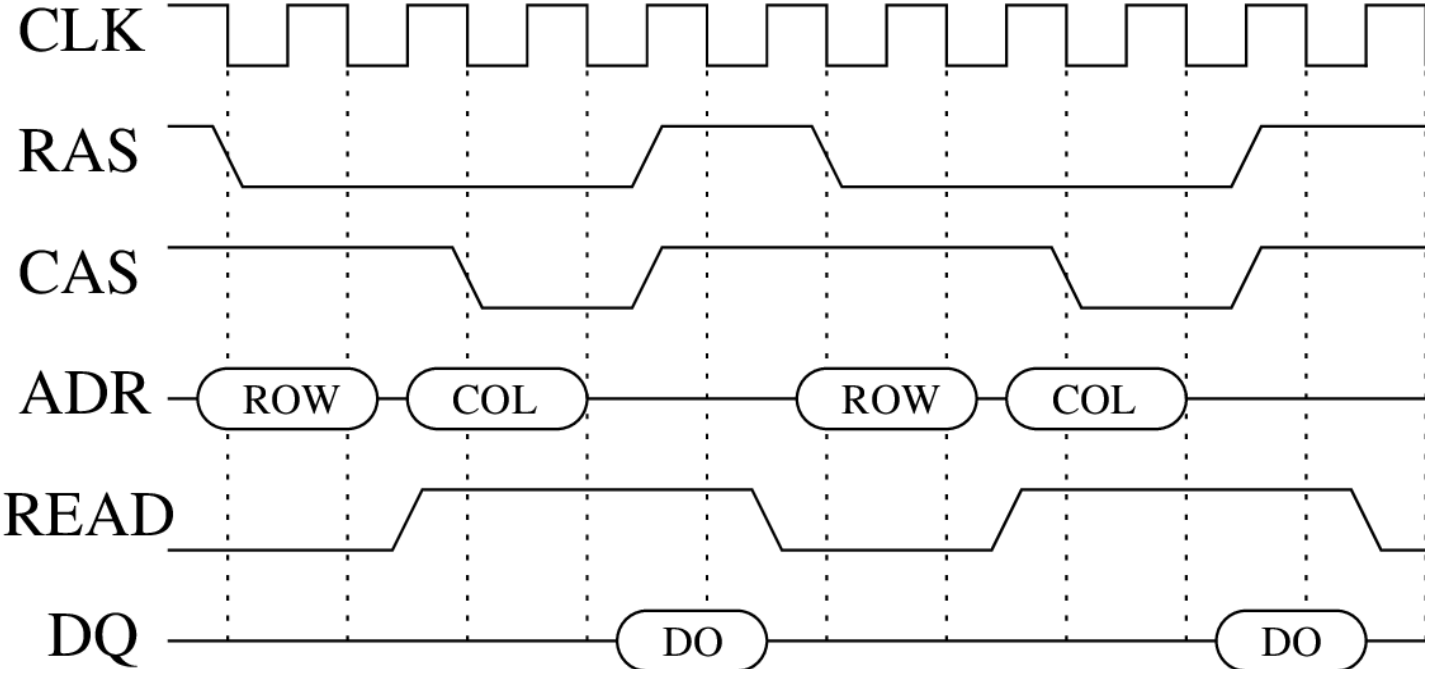
Now we know the design of DRAM. But how to write data to the memory? The time diagram of signal sequence is below. At first it is necessary to send to the memory address of the ROW and then the address of the COL. Together with the COL address data is written to the data bus. Memory has now one bus clock cycle to accept data from the bus. And the cycle repeats.



Read from DRAM

The reading from memory is similar to writing. The processor sends to the bus the sequence of the ROW and COL address signals and the memory chip writes data to the bus in the next bus clock cycle and the processor takes them.

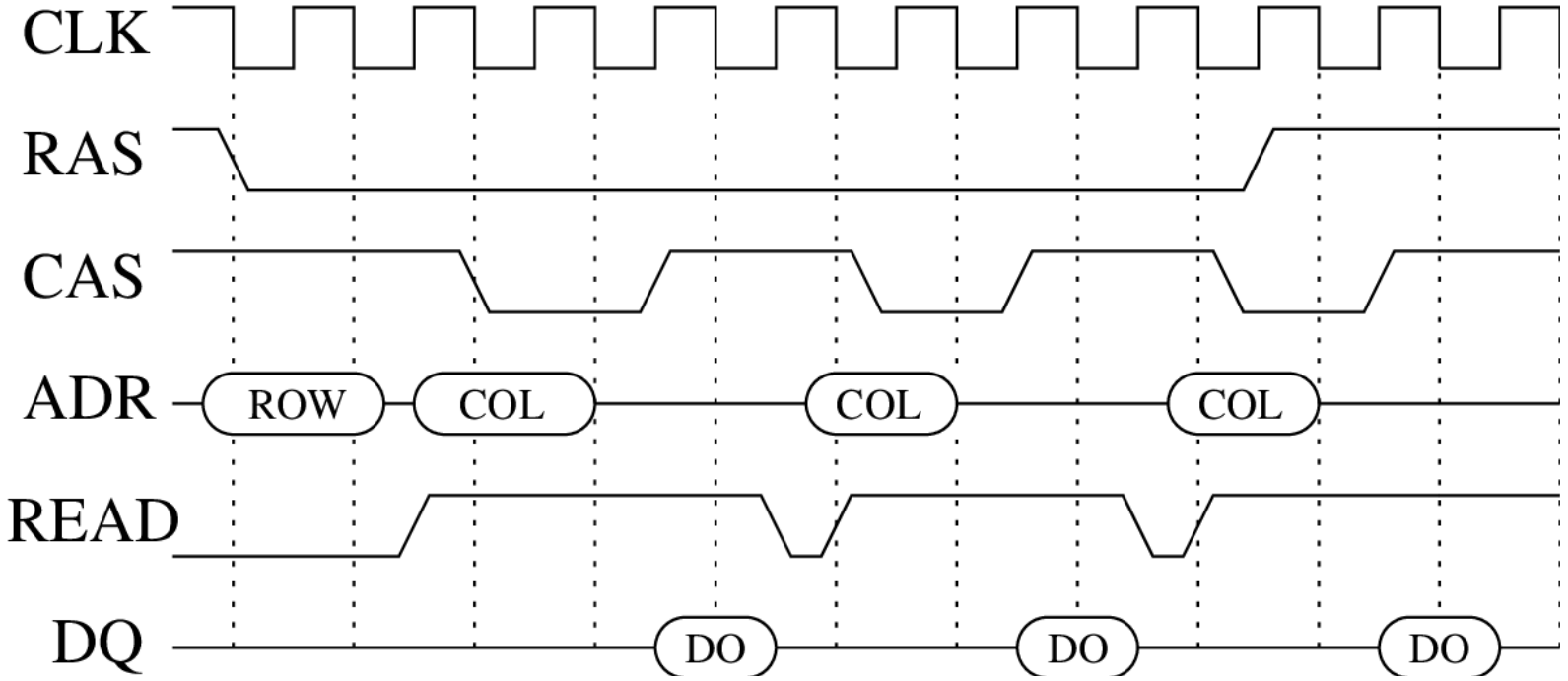
The reading and writing from the memory in this way is not too fast and on average it takes tens of nanoseconds.



Read from FP DRAM

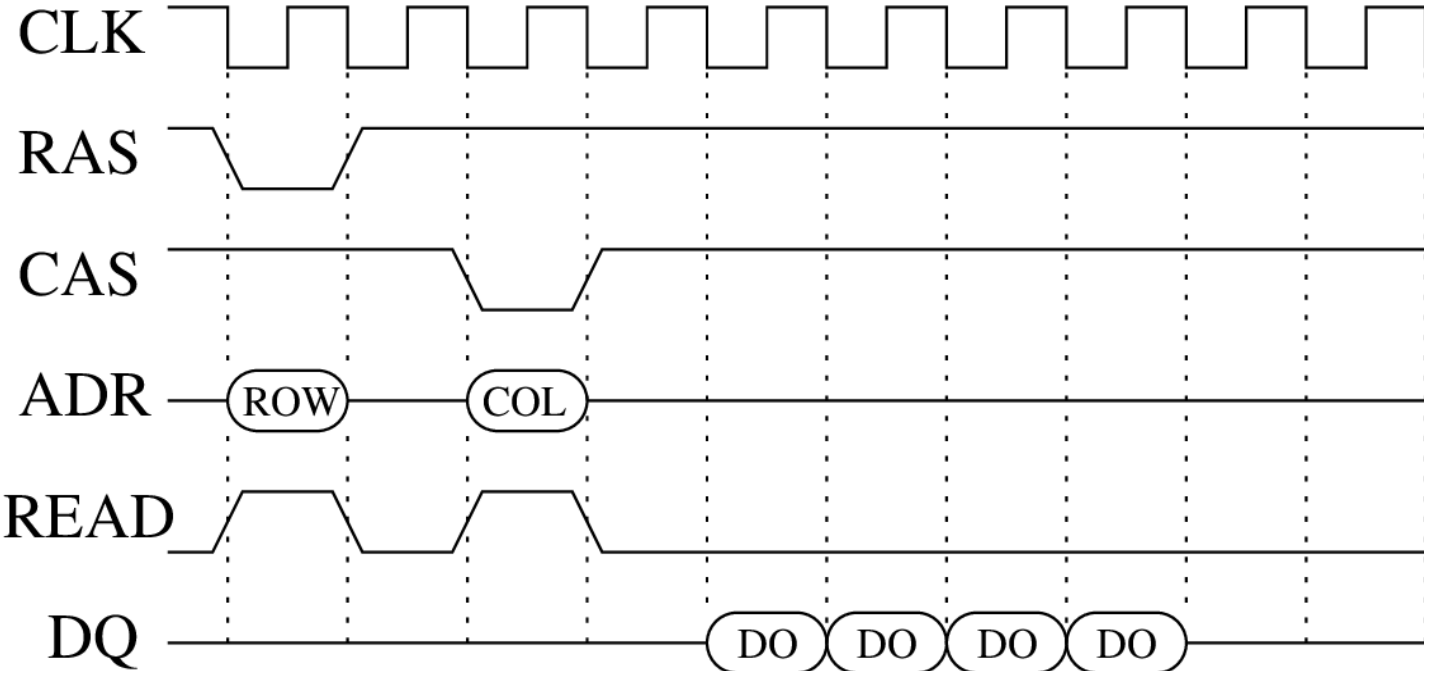
When the processor reads data from the memory, then the next reading is in the most cases from following addresses. Therefore better version of DRAM was designed – Fast Page DRAM.

When the processor reads data from following addresses, it is not necessary to send the ROW signal again. Only the following COL addresses are sent. Thus reading is faster.



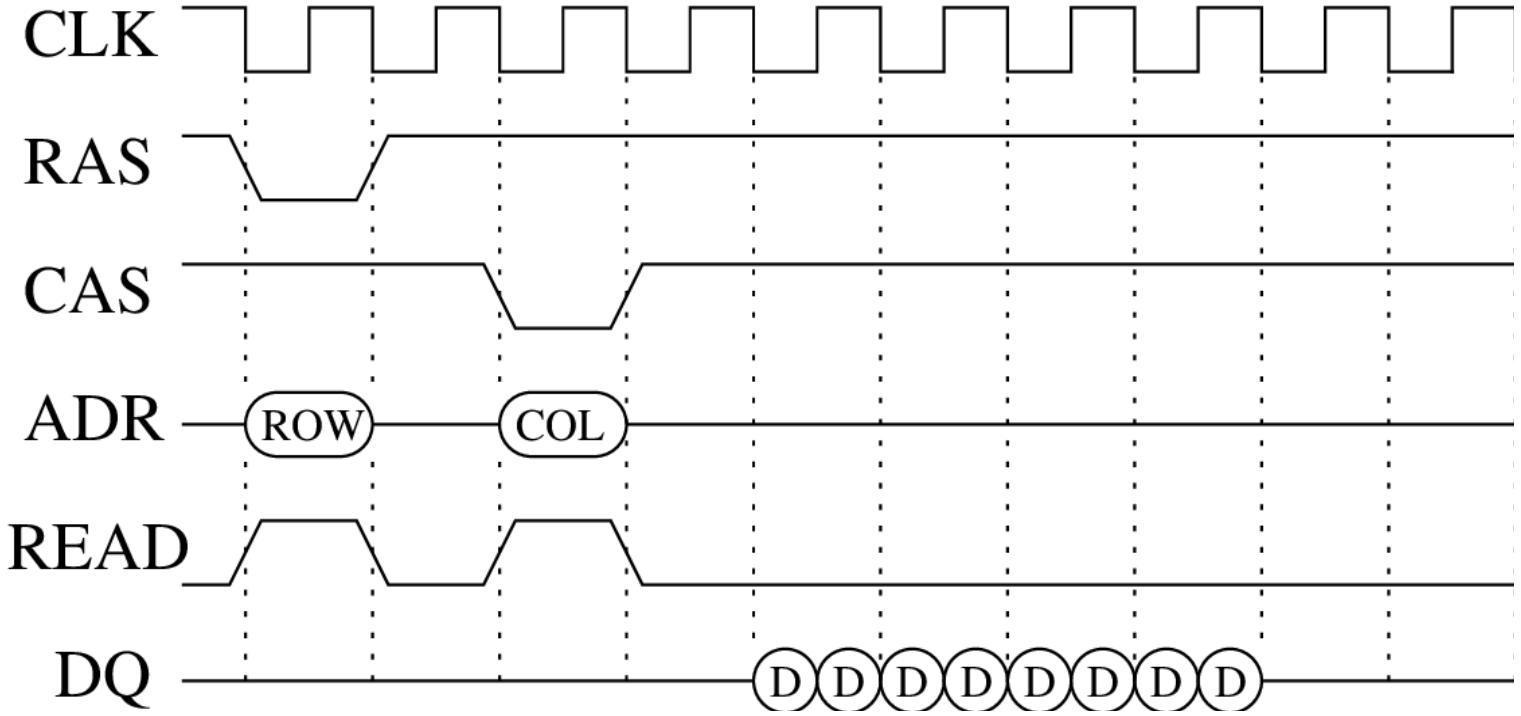
SDRAM

The speed of processors is increasing faster than speed of memories. Therefore their manufacturers introduced new version of memory to improve its performance – Synchronous DRAM. This memory obtains signals ROW and COL and then it generates the following COL signals automatically inside the chip and it sends data to the bus. No more control signals are needed. But the chip has to be synchronized with the processor clock signal.



DDR SDRAM

The latest generation of DRAM is DDR SDRAM technology – Double Data Rate SDRAM. This memory is two times faster than SDRAM. Data to the bus are sent two times per one clock cycle. This memory is even used in the computer as double channel memory and the final average time of reading is in units of nanoseconds. The maximum transfer speed is up to a few gigabytes per second.



DRAM Modules

DRAM memories were produced over the time in several versions:

- **DIP** – Dual In-line Package. As well as many other semiconductor chips, the DRAM memory was produced in this usual package too.
- **SIPP** – Single In-line Pin Packages. More DIP chips are mounted on module - simple circuit board - with pins on one edge. But the insertion of long line of pins to connector was complicated. Pins usually bent and installations were unreliable.
- **SIMM** – Single In-line Memory Module. The same as SIPP, but the connector is directly on the module edge without pins. First generation had 30 pins, second one 72 pins.
- **DIMM** – Dual In-line Memory Module. Designed for SDRAM a DDR SDRAM and for 64 bits data bus. It is successor of SIMM.
- **SO-DIMM** – Small Outline DIMM. It is designed for notebooks and embedded computers.

Static Memory

The static memory has all cells designed as the RS flip-flop. The information in cells is kept until the state is changed. It does not need refresh. Therefore these memories are called static.

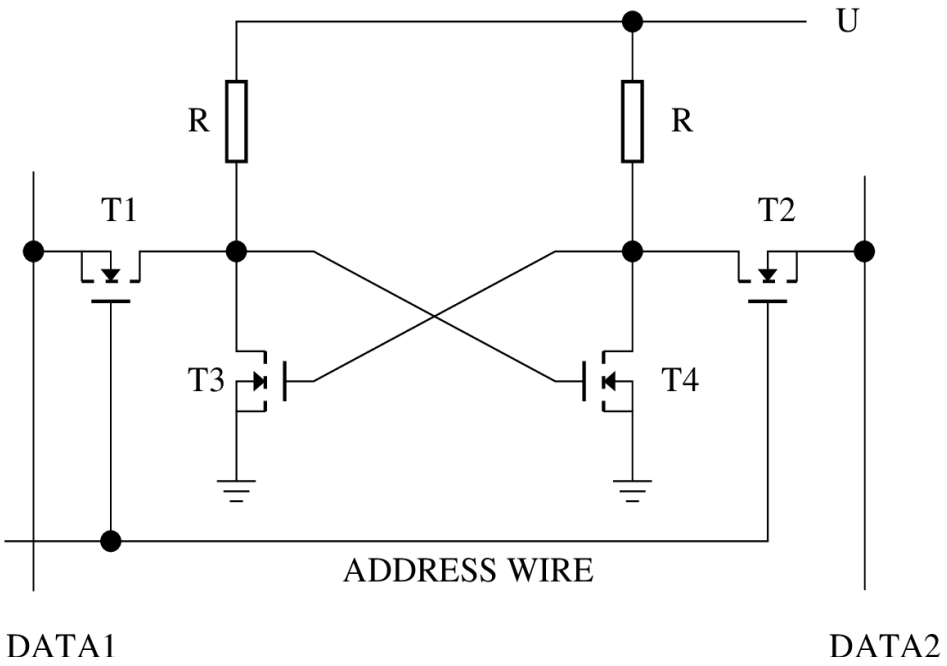
One bit can be saved in a cell consisting of four or six transistors. The scheme will be showed on the next slide.

In the cell implemented as four transistors version, two transistors work as flip-flop and use resistors as active load. Other two transistors are activated by the address wire and connect the cell to the data wire for the reading or writing.

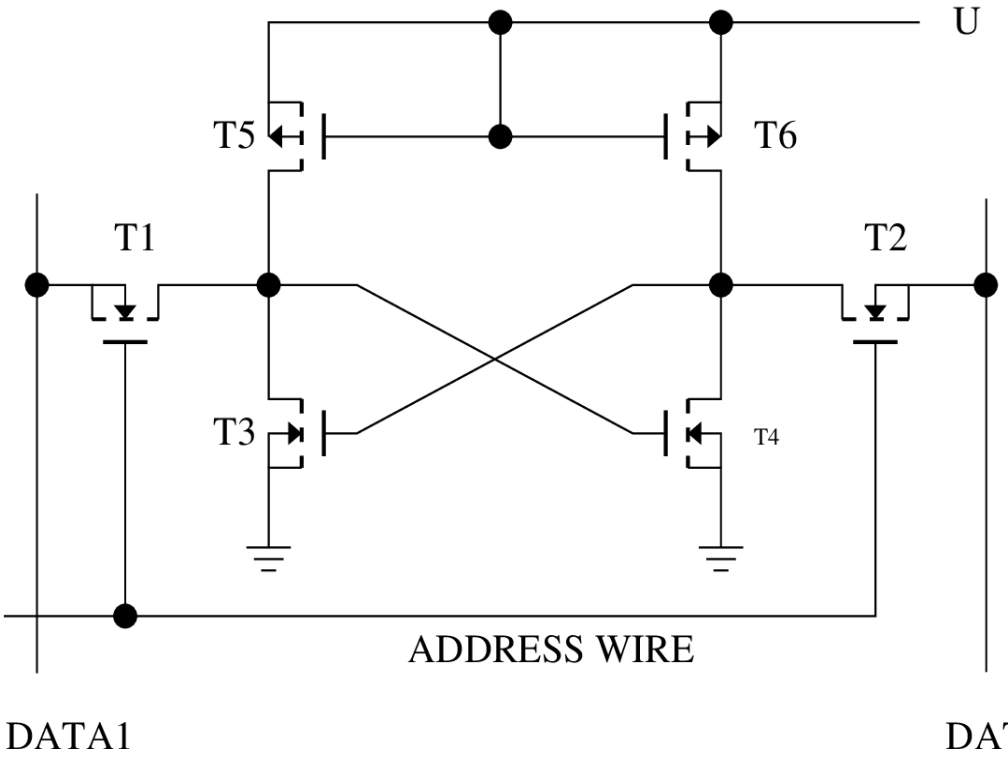
The more modern six transistors version of cell needs more transistors on the chip, but the resistor load is replaced by two transistors and the whole memory has lower power consumption. This design allows to design memories with greater capacity.

… Static Memory – Cells Scheme

Memory cell with 4 transistors



Memory cell with 6 transistors



... Static Memory

The flip-flop memory cell design allows to read and to write data very fast. Thus static memory is used for the fastest memory in computer. But overall, the high number of transistors does not allow to produce chips with the high capacity, as DRAM memory has.

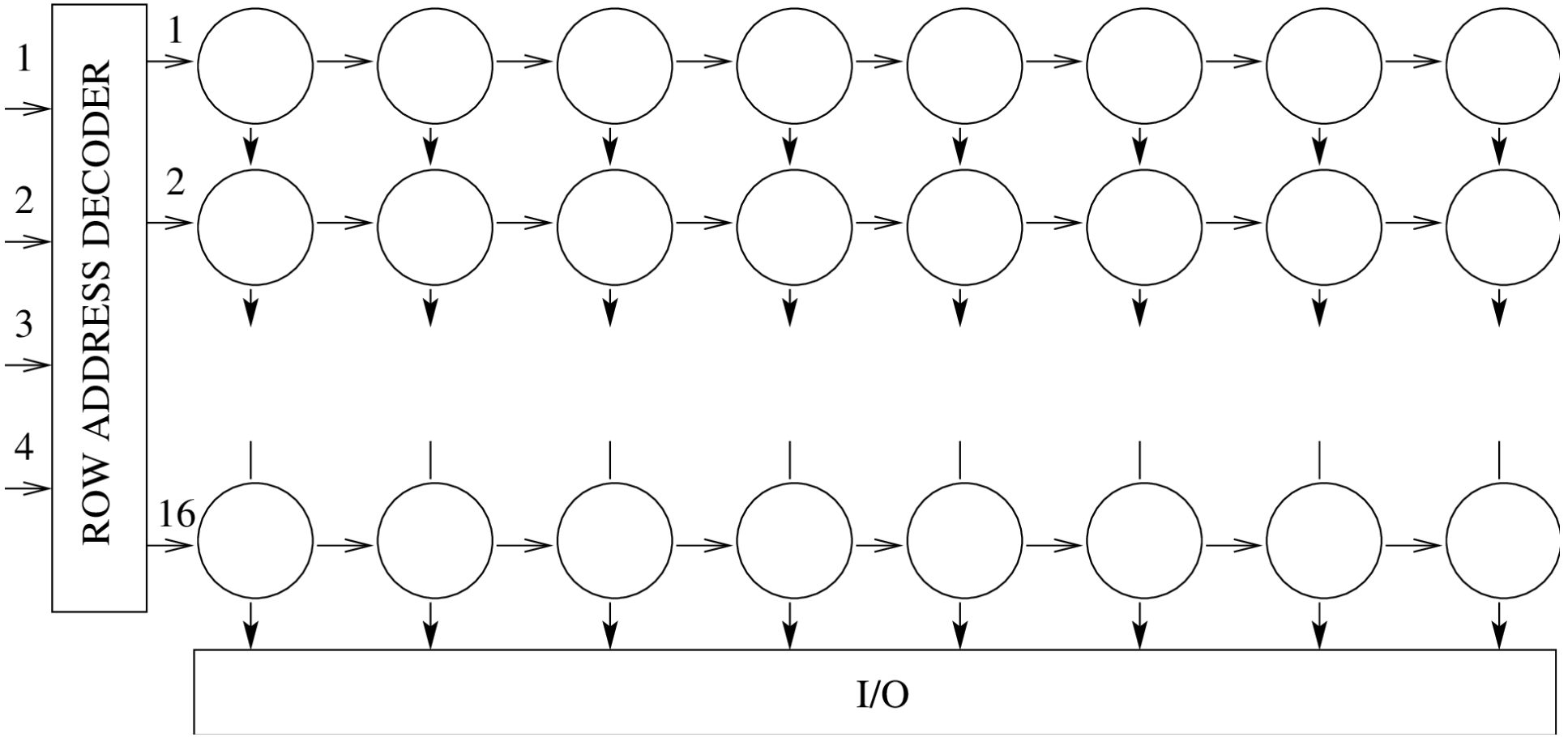
Because SRAM is used for memory with fast access, the internal structure is adapted accordingly. Memory cells are organized in the matrix, but all rows have their own full address. All cells in the row are activated in a single step.

Therefore it is necessary to have a more complex decoder in the chip and use more address wires. However with lower memory capacity its price is significantly higher.

The scheme of the internal organization of cells in the static memory is on the next slide.

… Static Memory – Internal Structure

The next scheme shows the chip with 4 bits address bus and 16 rows. Each row contains 8 bits.



Nonvolatile Memory

The SRAM and DRAM memories hold information only when they are powered by the electricity. But in all computers it is necessary to store a lot of information permanently – BIOS, firmware, measured data, configuration, etc. – in nonvolatile memories.

Nonvolatile memory needs power supply too, but only for data access. Information is held in special memory cells. They are constructed from the transistor with the insulated control electrode.

Several types of these memories are used in computers:

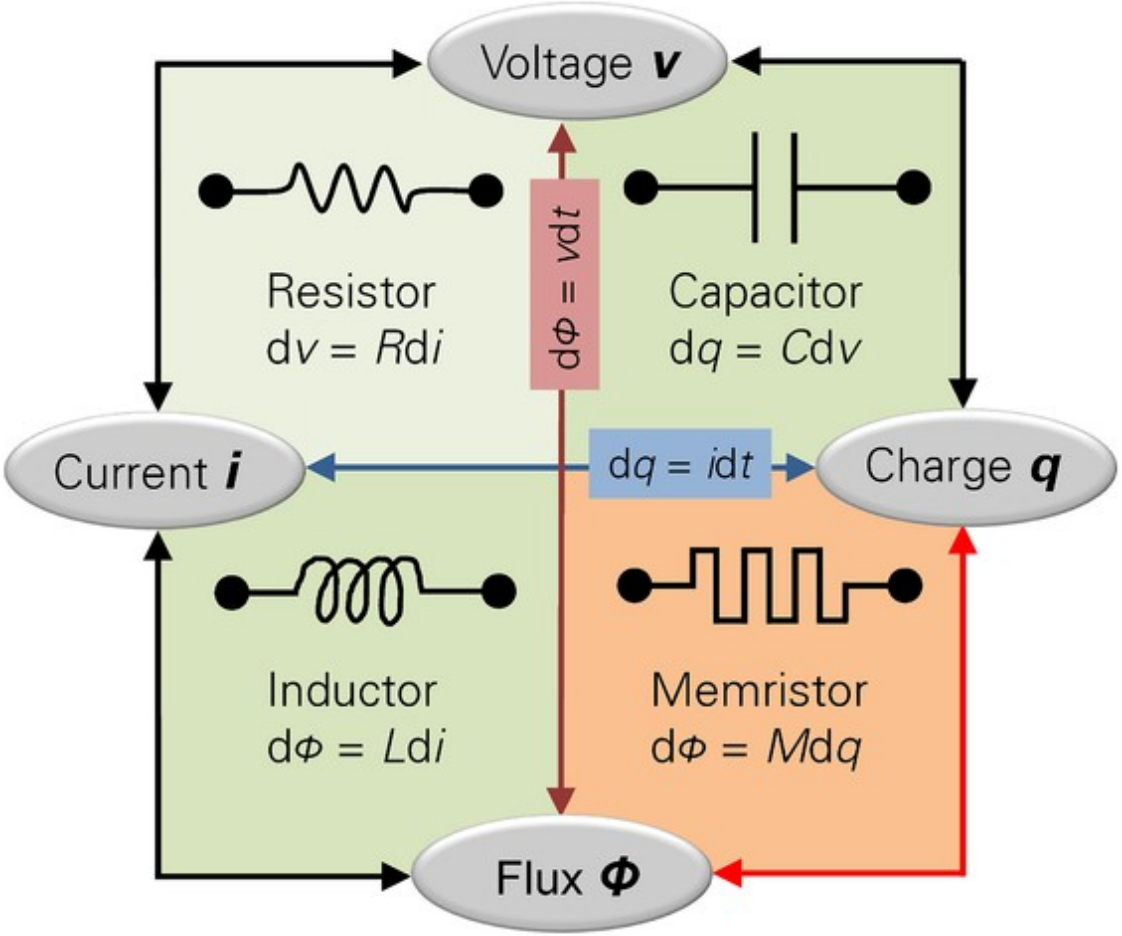
- **PROM or OTP** – Programmable Read Only Memory, nowadays called One Time Programmable. This memory is used where stored information will never change. For example in microcomputer using that memory.
- **EPROM** – Erasable PROM. This technology allows to clear the whole memory with ultraviolet light and program it again. But the chip has to be removed from board, which is inconvenient.

... Nonvolatile Memory

- **EEPROM** – Electrically Erasable PROM. This type of memory does not need UV light for erasing. It can be erased electrically. But for erasing it requires higher voltage, like the normal power supply. Normally the chip is supplied by 5V and for erasing 12V is used. It is unpleasant too that the memory must always be erased completely.
- **Flash** – the successor of EEPROM. Flash memory is now used in most computers as nonvolatile memory. It allows to erase only a small part of the memory and the modern version does not need higher voltage for the erasing. It works with normal voltage 5V or 3.3V. The disadvantages remain slow erasure and limited number of writes.
  Flash memory is not used only for programs in computers. Now it is more and more used as computer “disk” due to decreasing price. It is known as SSD (Solid State Disk) or USB Flash Drive.

Memristor

The fourth passive circuit component was already envisioned theoretically in 1971 but the first time it was made in 2008 by the Hewlett-Packard laboratories. The diagram with electrical properties:



... Memristor

The resistance of a memristor depends on the integral of the input applied to the terminals. Since the element "remembers" the amount of current that passed through it in the past, it was tagged with the name "memristor". Another way of describing a memristor is that it is any passive two-terminal circuit element that maintains a functional relationship between the time integral of current (called charge) and the time integral of voltage (often called flux, as it is related to magnetic flux). The slope of this function is called the memristance M.

Memristor is nano device that remember information permanently, switch it in nanoseconds, it is super dense and power efficient. This makes memristors potential replacements for DRAM, flash, and disk.

Probably in the near future, computers will use only one universal memory - “MRAM”. But redesign of computer architecture will be required.

Memory Testing

The memory testing is a complex problem. In most computers testing is performed off-line by special programs. Because the memory is internally organized into a 3D matrix, each cell can influence their own surrounding. Therefore testing programs have to combine many data patterns to detect error.

A well-known program for memory testing is available at:

http://www.memtest86.com

In servers, where off-line testing is impossible, the parity bit is used for error detection, or newer technology ECC – Error Correcting Code – to detect and recover small errors.

Operating system running on server obtains event from the memory controller and system administrator can prepare service action.

Cache Placement Policy

A Cache is a memory which holds the recently utilized data by the CPU. A block of memory cannot be placed randomly in the cache and may be restricted to a single cache line or a set of cache lines by the **cache placement policy**.

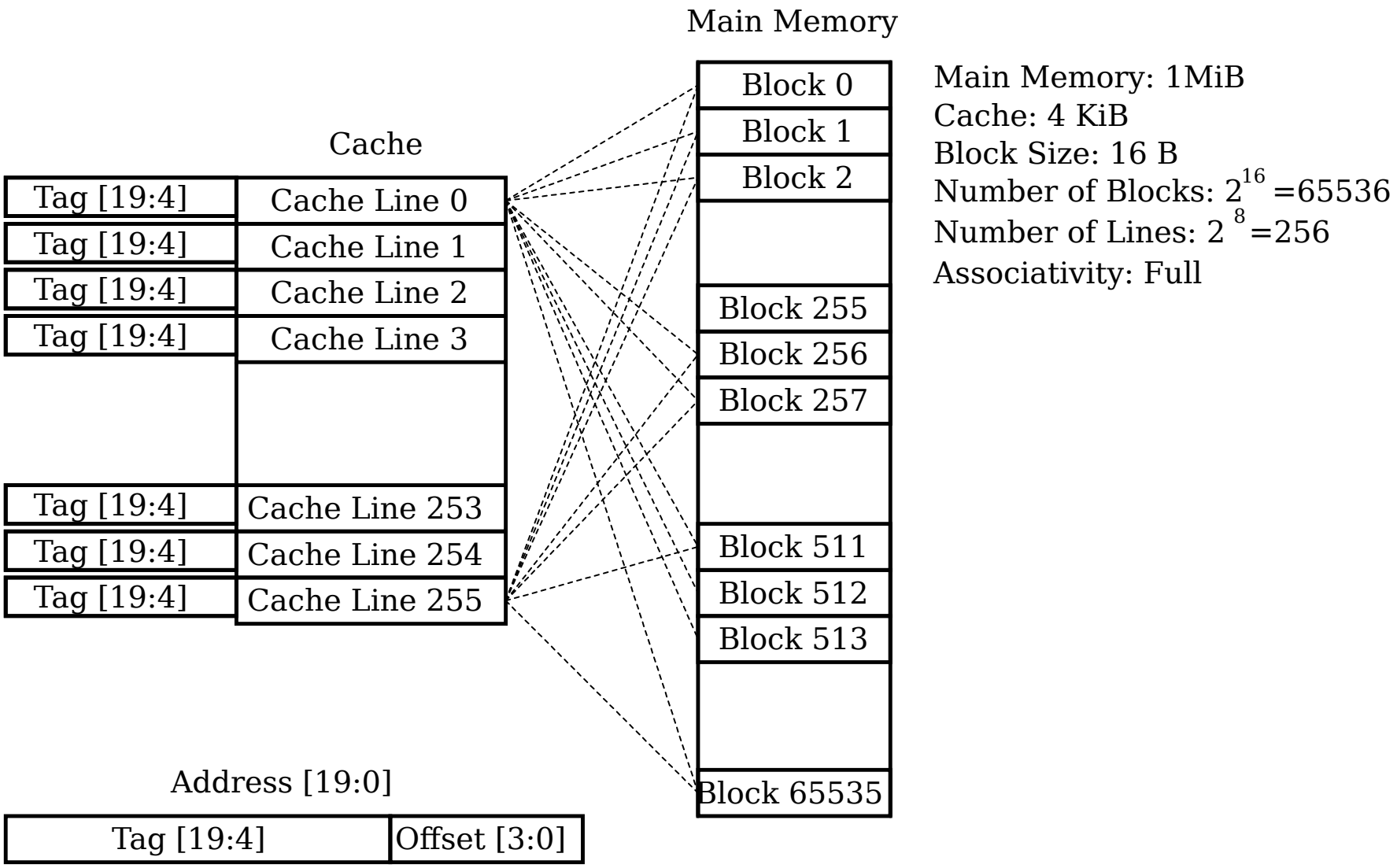
In the following text will be described three main placement policy:

- Fully associative cache,
- Direct-mapped cache,
- Set-associative cache.

In following examples will be described the mapping for:

- 1 MiB main memory,
- 4 KiB cache memory,
- block size 16 bytes.

Fully Associative Cache



Main Memory: 1MiB
Cache: 4 KiB
Block Size: 16 B
Number of Blocks: 2^{16} =65536
Number of Lines: 2^{8} =256
Associativity: Full

... Fully Associative Cache

In a fully associative cache, the cache is organized into a single cache set with multiple cache lines. A memory block can occupy any of the cache lines. The cache organization can be framed as matrix with 256 rows (in this example!).

To search a data in the cache the Tag field of the memory address is compared with tag bits associated with all the cache lines. If it matches, the block is present in the cache and is a cache hit. If it doesn't match, then it's a cache miss and has to be fetched from the lower memory. Based on the Offset, a byte is selected and returned to the processor.

In this example is 20-bit address splitted into two parts only. A four bits are **Offset** for addressing inside one block and 16 bites are for a **Tag**.

... Fully Associative Cache

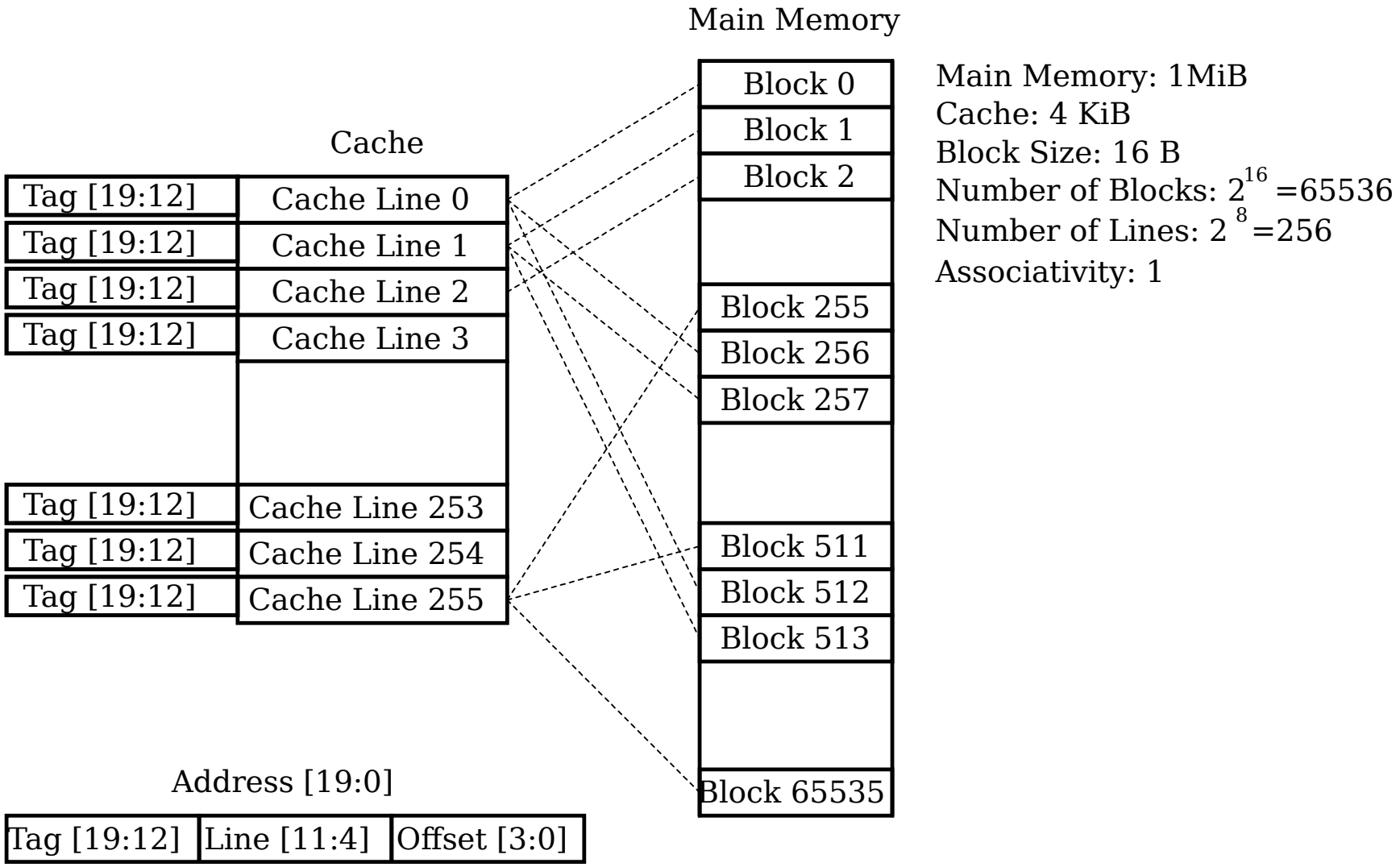
Advantages:

- Fully associative cache structure provides us the flexibility of placing memory block in any of the cache lines and hence full utilization of the cache.
- The placement policy provides better cache hit rate.
- It offers the flexibility of utilizing a wide variety of replacement algorithms if a cache miss occurs

Disadvantages:

- The placement policy is slow as it takes time to iterate through all the lines.
- The placement policy is power hungry as it has to iterate over entire cache set to locate a block.
- The most expensive of all methods, due to the high cost of associative-comparison hardware.

Direct-Mapped Cache



... Direct-Mapped Cache

In a direct-mapped cache structure, the cache is organized into multiple sets with a single cache line per set. Based on the address of the memory block, it can only occupy a single cache line. The cache can be framed as a matrix with 256 columns (in this example!).

The set is identified by the line bits of the address. The tag bits derived from the memory block address are compared with the tag bits associated with the set. If the tag matches, then there is a cache hit and the cache block is returned to the processor. Else there is a cache miss and the memory block is fetched from the main memory.

In this example is the address splitted into three parts: 4 bits are for a block **Offset**, 8 bits is for a **Line** selection and 8 bit is a **Tag**.

... Direct-Mapped Cache

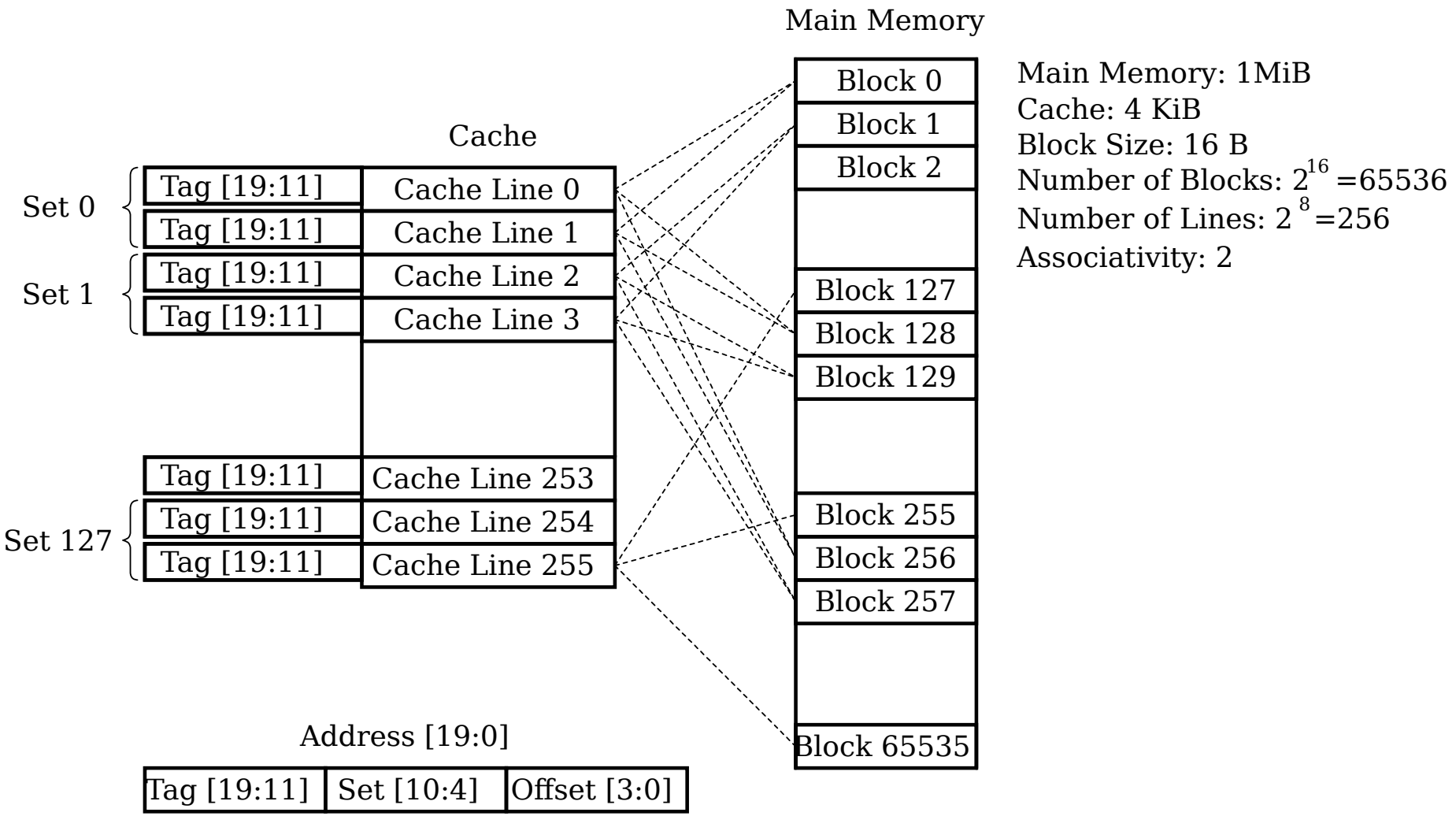
Advantages:

- This placement policy is power efficient as it avoids the search through all the cache lines.
- The placement policy and the replacement policy is simple.
- It requires cheap hardware as only one tag needs to be checked at a time.

Disadvantages:

- It has lower cache hit rate, as there is only one cache line available in a set. Every time a new memory is referenced to the same set, the cache line is replaced, which causes conflict miss.

Set-Associative Cache



Main Memory: 1MiB
Cache: 4 KiB
Block Size: 16 B
Number of Blocks: 2^{16} =65536
Number of Lines: 2^{8} =256
Associativity: 2

... Set-Associative Cache

Set-associative cache is a trade-off between direct-mapped cache and fully associative cache. A set-associative cache can be imagined as a matrix with size 127x2 (only in this examle!). The cache is divided into 127 sets and each set contains 2 cache lines. A memory block is first mapped onto a set and then placed into any cache line of the set.

Many processor caches in today's designs are either direct-mapped, two-way set-associative, or four-way set-associative.

In this example is address splitted into three parts: 4 bits are a **Offset** inside block, 7 bits are a **Set** and 9 bits are a **Tag**.

... Set-Associative Cache

Advantages:

- The placement policy is a trade-off between direct-mapped and fully associative cache.
- It offers the flexibility of using replacement algorithms if a cache miss occurs.

Disadvantages:

- The placement policy will not effectively use all the available cache lines in the cache and suffers from conflict miss.

VSB TECHNICAL UNIVERSITY OF OSTRAVA
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
DEPARTMENT OF COMPUTER SCIENCE