### 3.5.2 Main Memory

Main memory is the next level of the hierarchy, downstream from the caches. Requests to load and store data are initiated by the Memory Controller Unit (MCU). In the past, this circuit was located inside the motherboard chipset, in the north bridge chip. But nowadays, most processors have this component embedded, so the CPU has a dedicated memory bus connecting it to the main memory.

Main memory uses DRAM (Dynamic Random Access Memory), technology that supports large capacities at reasonable cost points. When comparing DRAM modules, people usually look at memory density and memory speed, besides its price, of course. Memory density defines how much memory the module has, measured in GB. Obviously the more available memory the better as it is a precious resource used by the OS and applications.

Performance of main memory is described by latency and bandwidth. Memory latency is the time elapsed between the memory access request is issued and when the data is available to use by CPU. Memory bandwidth defines how many bytes can be fetch per some period of time, usually measured in gigabytes per second.

DDR (double data rate) DRAM technology is the predominant DRAM technology supported by most CPUs. Historically, DRAM bandwidths have improved every generation while the DRAM latencies have stayed the same or even increased. The table 2 shows the top data rate, peak bandwidth, and the corresponding reading latency for the last three generations of DDR technologies. The data rate is measured as a million transfers per sec (MT/s). The latencies shown in this table correspond to the latency in the DRAM device itself. Typically, the latencies as seen from the CPU pipeline (cache miss on a load to use) are higher (in the 50ns-150ns range) due to additional latencies and queuing delays incurred in the cache controllers, memory controllers, and on-die interconnects. See an example of measuring observed memory latency and bandiwdth in section 4.10.

Table 2: Performance characteristics for the last three generations of DDR technologies.

| DDR Generation | Year | Highest Data Rate(MT/s) | Peak Bandwidth (Gbytes/s) | In-device Read Latency(ns) |
|---|---|---|---|---|
| DDR3 | 2007 | 2133 | 12.8 | 10.3 |
| DDR4 | 2014 | 3200 | 25.6 | 12.5 |
| DDR5 | 2020 | 6400 | 51.2 | 14 |

It is worth to mention that DRAM chips require memory cells being periodically refreshed. Because the bit value is stored as the presence of an electric charge on a tiny capacitor, it can lose its charge as the time passes. To prevent this, there is a special circuitry that reads each cell and writes it back, effectively restoring the capacitor's charge. While a DRAM chip is in its refresh procedure, it is not serving memory access requests.

DRAM module is organized as sets of DRAM chips. Memory *rank* is a term that describes how many sets of DRAM chips exist on a module. For example, a single-rank (1R) memory module contains one set of DRAM chips. A dual-rank (2R) memory module has two sets of DRAM chips, therefore doubling the capacity of a single-rank module. Likewise, there are quad-rank (4R) and octa-rank (8R) memory modules available for purchase.

Each set of memory chips consists of multiple chips. Memory *width* defines the width of the bus of each chip in a set and consequently, the number of chips in a set. Memory width can be one of three values: `x4`, `x8` or `x16`, which define how wide is the bus that goes to each chip. As an example, figure 12 shows the organization of 2Rx16 dual-rank DRAM DDR4 module, total 2GB capacity. There are four chips in each set, with a 16-bit wide bus. Combined, the four chips provide 64-bit output. The two ranks are selected one at a time through a chip set select signal.

There is no direct answer whether performance of single-rank or dual-rank is better as it depends on the type of application. Switching from one rank to another through chip select signal needs additional clock cycles, which may increase the access latency. On the other hand, if a rank is not accessed, it can go through its refresh cycles in parallel while other ranks are busy. As soon as the previous rank completes data transmission, the next rank can immediately start its transmission. Also, single-rank modules produce less heat and are less likely to fail.

Going further, we can install multiple DRAM modules in a system to not only increase memory capacity, but also memory bandwidth. Setups with multiple memory channels are used to scale up the communication speed between the memory controller and the DRAM.

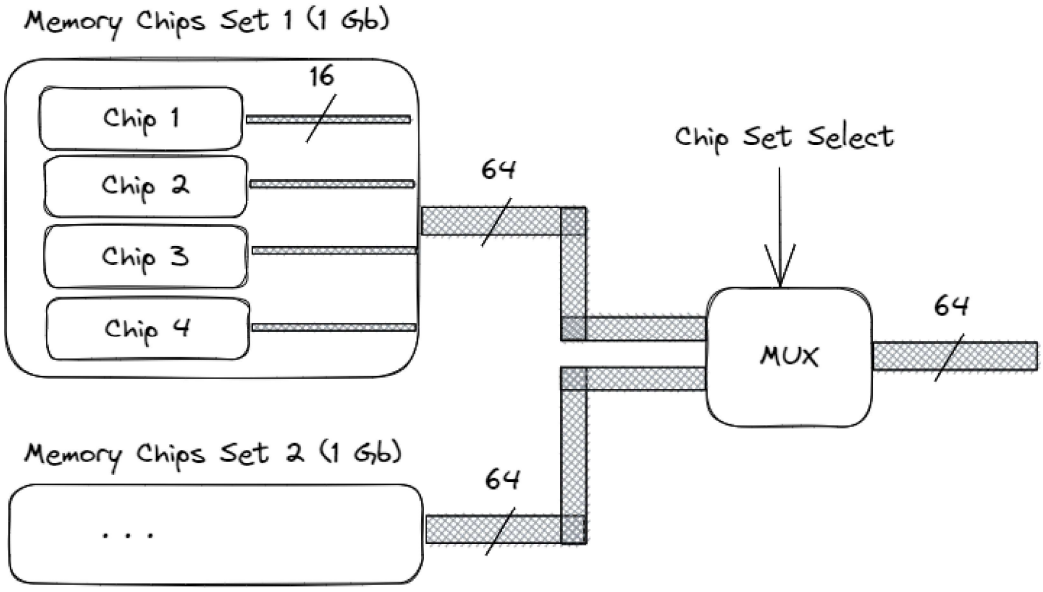


Figure 12: Organization of 2Rx16 dual-rank DRAM DDR4 module, total 2GB capacity.

A system with a single memory channel has a 64-bit wide data bus between the DRAM and memory controller. The multi-channel architectures increase the width of the memory bus, allowing DRAM modules to be accessed simultaneously. For example, the dual-channel architecture expands the width of the memory data bus from 64 bit to 128 bit, doubling the available bandwidth, see figure 13. Notice, that each memory module, is still a 64-bit device, but we connect them differently. It is very typical nowadays for server machines to have four and eight memory channels.

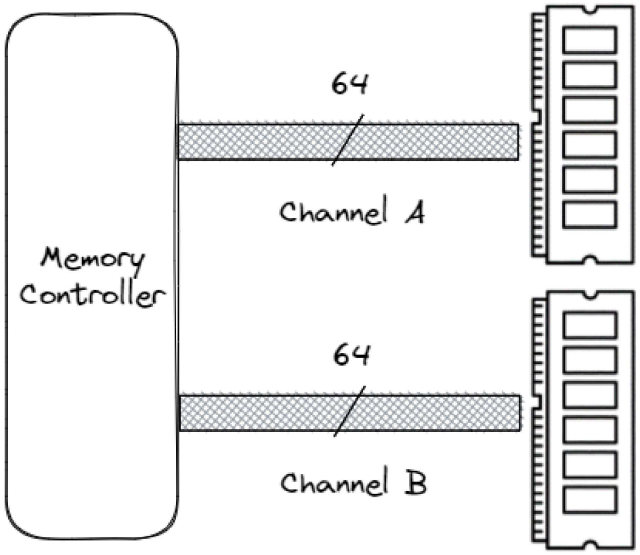


Figure 13: Organization of a dual-channel DRAM setup.

Alternatively, you could also encounter setups with duplicated memory controllers. For example, a processor may have two integrated memory controllers, each of them capable of supporting several memory channels. The two controllers are independent and only view their own slice of the total physical memory address space.

We can do a quick calculations to determine the maximum memory bandwidth for a given memory technology, using

a simple formula below:

$$\text{Max. Memory Bandwidth} = \text{Data Rate} \times \text{Bytes per cycle}$$

For example, for a single-channel DDR4 configuration, the data rate is `2400 MT/s` and 64 bits or 8 bytes can be transfered each memory cycle, thus the maximum bandiwdth equals to `2400 * 8 = 19.2 GB/s`. Dual-channel or dual memory controller setups double the bandwidth to `38.4 GB/s`. Remember though, that those numbers are theoretical maximums, that assume that a data transfer will occur at each memory clock cycle, which in fact never happens in practice. So, when measuring actual memory speed, you will always see a value lower than the maximum theoretical transfer bandwidth.

In order to enable multi-channel configuration, you need to have a CPU and a motherboard that supports such architecture and install an even number of identical memory modules in the correct memory slots on the motherboard. The quickest way to check the setup is by running a hardware identification utility like `CPU-Z` or `HwInfo`. But also, you can run the memory bandwidth benchmarks like Intel `mlc` or `Stream`.

While increased memory bandwidth is generally good, it does not always translate into increased system performance and is highly dependent on the application.