Development of memory controller for today’s Elbrus microprocessors

I. A. Petrov

1 MCST JSC, Moscow, Russia

The introduction of a new generation of microprocessors that belong to the Elbrus family and involve the introduction of a network-on-chip, requires the development of efficient means of access to DDR random access memory channels for network nodes. The paper includes a solution to this issue related to the interaction between DDR4 RAM and Elbrus-16СВ, 16-core microprocessor, which demands higher standards of an available capacity and peak bandwidth of memory channels. When designing Elbrus-16СВ microprocessor, higher energy efficiency and reliability are also between main objectives. When performing the tasks set, an important component was adaptation of the memory controller, successfully applied in the microprocessors produced by MCST JSC, to DDR4 3DS standard compliance, taken as a basis for the use in a number of recent developments. It provides a four-time higher available RAM capacity without a directly proportional growth of energy consumption. The paper includes a structure of the memory controller and made decisions. These make it possible to increase the target frequency in operations of the device by 30% up to 800 MHz and increase operation reliability of the memory channel.

Keywords: network-on-chip, memory controller, DDR4 3DS, design

Introduction

The constant increase in the number of blocks in the microprocessor chip has become one of the defining problems of modern microelectronics. This is due to the fundamental development of modern computer architectures (increasing the number of processor cores in a chip or the distribution of last-level cache banks over a chip), as well as to the expansion of the practical use of computing systems, which leads to an increase in the external field interacting with a microprocessor. In this regard, the most important design factor is the choice of the communication medium of the crystal that connects the blocks, which would have the necessary scalability, performance, and flexibility. A modern solution to this problem is network-on-chip (NoC), which is the most appropriate architecture of the microprocessor being developed in one of its implementations [1].

One of the most important tasks in the design of the first microprocessor of the Elbrus family with NoC technology is the development of nodes that provide network interaction. Fundamentally, the functions performed by each of them are divided into two groups. The functions of the first group provide the reception and transmission of network packets with the provided quality of services (router functions). The functions of the second group provide effective communication with the microprocessor architectural unit connected to the node. In this sense, the interaction of microprocessor nodes with modules of the RAM subsystem is of particular importance, which has a decisive impact on the final performance of the microprocessor. The purpose of the article is to consider this problem.
According to the characteristic industrial developments, the tendency to use substantially developed solutions of previous projects in the newly created products whenever possible, the results described in the article were obtained on the basis of the fact that when organizing access to the RAM, the authors used a controller of own design [2], which evolved to support the DDR4 standard (JEDEC Standard DDR4 SDRAM, JESD79–4B) in the «Elbrus-8SV» microprocessor. The article provides an overview of the improvements and modifications of this device, which allow obtaining the indicators set in the Elbrus-16SV project [3], and will be used in a series of planned projects. In relation to this, increased requirements are considered for the available volume and peak performance of the RAM, as well as for the functions of the RAS category (reliability, availability, maintainability).

Random-access memory of the Elbrus-16SV microprocessor

**DDR43DS standard support**

A microchip made according to the DDR43DS standard [4], which provides for a significantly increased RAM volume as compared to previous developments of MCST JSC, is a vertical assembly (stack) of conventional SDRAM DDR4 chips assembled in one housing, built on the «communication through silicon» technology (Through silicon via, TSV). The 3DS protocol introduces the concept of «logical rank» (one chip per stack) in addition to the physical rank. The maximum number of logical ranks in the stack is eight; therefore, compared with the capabilities of the previous Elbrus-8SV microprocessor, the maximum amount supported by the memory controller in one channel increases by eight times up to 1 TB. At the moment, the main manufacturers of memory chips release 128 GB DDR43DS4H Stack modules, which makes it possible to accumulate 2 TB memory for one Elbrus-16SV processor when two slots are filled in each channel.

**Interleaving mechanism**

In each logical bank (size ranging from 256 MB to 2 GB), one can use only one page of memory (8 KB) at a time. Switching to another page within one logical bank is possible in 45–50 ns. To increase the number of simultaneously used pages in the memory channel, the interleaving mechanism is used – the process of mixing the initially consecutive pages of memory between logical banks located in different physical and logical ranks and groups of logical ranks. At the stage of system initialization, the positions of the address bits of the request are set in the memory controller, which determine the number of the logical bank in the space of logical ranks where the request should go. The use of inter-stack interleaving to a greater extent than interleaving between physical ranks can increase the real channel throughput due to lower delays in switching between stacks.

Increasing the capacity of the memory module due to the use of DDR43DS does not lead to a proportional increase in power consumption and a decrease in the operating frequency, since the load on the data bus remains comparable to conventional DDR modules and does not depend on the number of logical ranks in the stack [5].

**Modules with ECC correction support**

When using modules with ECC correction support, before starting the system with the RAM, it will be necessary to fill in the ECC code in it with the correct data, since after turning on the power, the contents of DRAM cells with ECC code do not correspond to the contents of DRAM cells with an attached data block [6]. Full initial memory initialization can be done programmatically, but in order to reduce execution time, the memory controller...
has the built-in opportunity of cleaning the RAM with hardware. This process, for example for 32 GB of the DDR4–2400 module, takes approximately 2 seconds. In connection with the use of more voluminous 3DS-chips, the memory hardware cleaning algorithm has been finalized. The memory controller no longer blocks the execution of system requests until the end of cleaning. It runs in the background, and the boot program of the system is able to work with a piece of memory immediately after it has been initialized.

**System solutions**

RAM channels controlled by independent devices are located on the network-on-chip nodes. A simplified block diagram of one such node is shown in Fig. 2.

In this implementation, the memory controller (MC) and the physical layer DDR for the first time in a series of Elbrus processors are combined structurally and spatially into one unit. This integration has the following advantages:

- reducing the time for physical design and the ability to perform it at no additional cost in parallel or subsequent developments of Elbrus processors with the introduction of on-chip networks;
- simplified mem_clk clock signal grid construction due to the localization of devices using it. At the same time, there is no need to use a synchronization signal transmission with data technique called Source-Synchronous Clocking. It was used in Elbrus-8SV due to long-distance transfers between the components of the access device in the RAM and intersections along the transmission path of independently controlled VDD power domains. In this embodiment, the clock becomes asynchronous with respect to the original mem_clk. This requires the insertion of an additional level of resynchronization, which increases the delay in data transmission.

The power system of the DDR4 physical layer is sensitive to asynchronous noise; therefore, an independent VDD power domain is allocated for the access device in the RAM. The transition across the border of the power domains is carried out by half-clock register transfers with the prohibition of the insertion of intermediate combinational logic.

**Controller functionality**

**Controller structure**

The main components of the controller are presented in Fig. 3.

The write data buffer is located in the controller, while the read data buffer, to which data received from the memory channel is sent from the physical layer DDR4 (see Fig. 2), does not directly enter the controller. Both buffers are filled in accordance with the logic defined by the scheduler based on the current status of the query registry.

A feature of the DDR4 interface is the need to observe delays – protocol locks between operations on the memory bus, the value of which depends on the time taken by related operations. Lock control is performed independently for each logical bank using the same type of timer counters. The appearance of the concept of logical rank is equivalent to adding a fourth dimension to the existing three-dimensional structure of the array of logical banks: physical rank – group of logical banks – logical bank within the group. In this regard, an extensive expansion of the number of timer counters was performed in the protocol lock controller. A similar expansion was made in the energy-saving and memory regeneration controller,

Figure 2. Simplified structure of RAM channel access point
since it is impossible to simultaneously issue a command for memory regeneration in several logical ranks.

**Scheduler and application register**

In the memory controller, the scheduler is combined with the registry, which performs the arbitration of applications (selection of the next application for execution). The selection is based on the parameters stored in each cell of the registry. These include both system parameters (address, type of operation), and parameters modified by the memory controller (the sign of data readiness for the write operation, the age of the request, the stage of request processing). The task of the scheduler is to change the order of processing applications in such a way as to minimize the average time spent on one application by reducing the number of auxiliary operations of opening and pre-loading pages, as well as eliminating the alternation of read and write operations. The principle of the scheduler is to ensure the maximum load of the DDR4 interface with memory modules, i.e., to eliminate empty cycles on the data bus. The internal organization of the scheduler is a system of series-connected filters, at the output of which there is a unit for the formation of elementary operations for the address-command bus of the DDR4 interface. The following filters are used:

- a resource filter that delays requests that cannot be completed due to insufficient resources (if there are no places in the reading option to save the read data in the read data buffer, and there is data in the write data buffer in the case of the write option);
- filter of the address dependency, which ensures the correctness of the sequence of calls to one address and identifies conflicts of the WAR (write after reading) and RAW (read after write) types, prohibiting the execution of requests that come later;
- priority filter for executing a request to an open page of a logical bank, excluding additional operations of opening pages. The request is screened out if it falls into the closed page, although at the same time there are requests addressed to the open page;
- a priority filter for read operations, aimed at delaying write requests when there are read requests due to the fact that read operations are more critical in terms of overall system performance;
- a protocol lock filter, based on the status of the protocol lock controller timers, it delays requests that are currently violating the DDR4 protocol;
- an age filter that skips the oldest request from all available at the input.

The increase in the number of cores and the supported amount of memory requires providing a sufficient number of positions in the application register. Together with the envisaged increase in the peak bandwidth of the memory channel from 19.2 GB/s ($f_{mem\_clk} = 600$ MHz) to 25.6 GB/s ($f_{mem\_clk} = 800$ MHz), this would lead, first, to an unacceptably long execution time of the physical memory controller synthesis, second, to the increase in time it takes for queries to pass through the filter system from one target mem\_clk clock cycle to several, which would complicate the filter system due to the need for pipelining.

In this regard, it was determined that the problems with the synthesis of the device are due to a sharp increase in the combinational logic in the address dependency filter and the priority filter for executing a request to the open page of the logical bank when using address comparators in the «each cell with each» format. As a result, the array of request addresses was in fact an associative memory (content addressable memory, CAM) with N (number of cells) search ports. To avoid this poorly scalable solution, the following features were added for each request to the parameters modified by the memory controller:

1) «there is a request in the registry at the intersecting address with a lesser age»;
2) «the logical bank for this request is busy».

Both features are updated based on address comparison when new requests arrive from two input ports and after the scheduler issues requests to the DDR4 interface (two ports). Thus, the number of CAM memory ports of the address array decreased to four.

![Controller’s functional structure](image-url)
allowed converting the dependence of the number of necessary elementary logical elements on the increase in the number of cells in the registry from quadratic to linear. Accordingly, the physical synthesis of the device began to be carried out in an acceptable time, and the passage of requests through the filter system remained single-cycle. As a result, the number of cells in the application registry was doubled relative to this indicator of the Elbrus-8SV processor memory controller, while maintaining the target operating frequency of the 800 MHz memory controller (DDR4–3200).

The RAS factor
In the new version of the memory controller, much attention was paid to improving the functionality of RAS (reliability, availability, serviceability) [7] due to the fact that the likelihood of hardware failures in the memory chips and in transfers via the DDR4 interface increases with an increase in the target frequency and available volume.

The mechanics for fixing the following two types of errors on the DDR4 interface are added: parity errors on the address-command bus and CRC (cyclic redundancy code) errors for write data. When an error is detected, the controller re-issues a sequence of commands to the DDR4 interface, the execution of which is not guaranteed in the memory module in the temporary vicinity of the failure instead of issuing a signal about an abnormal system shutdown, while ensuring memory coherence.

To control the integrity of the RAM, a patrol mechanism is implemented — periodic (software-tuned) reading through read-modify-write operations of the entire memory in order to correct the accumulated single ECC error. According to the practical experience of previous projects, when a multiple ECC error occurred, it was possible to expand the registration of the place of its occurrence up to the SDRAM chip number and the data bus bit, thereby, for example, reducing the time required to detect a persistently faulty data line in the area from the memory module to the processor.

The memory controller has the ability to configure the level of the reference voltage of the data bus in the memory chips, necessary to obtain eye diagrams characterizing the operability of the memory channel [8], on the fly.

Conclusions
The selected switching system in the developed Elbrus-16SV microprocessor required a reorganization of the RAM access subsystem. As a result, a new structural block was formed — the network-on-chip access node to the RAM. The universal nature of the decision makes it possible to quickly implement it in a number of further developments of the company. In the course of the work, the task of adapting a memory controller of the authors’ own design to the increased requirements for the volume, bandwidth capacity and reliability of memory channels was also solved. The increase in the volume of RAM is provided by the introduction of support for the DDR4 3DS standard. The maximum frequency of the memory controller was brought up to 800 MHz, which was made possible due to a significant reduction in equipment for the filter system of the controller scheduler. At the same time, the size of the registry of applications was doubled, which contributes to the optimal use of the memory channel bandwidth.

REFERENCES
1. Nedbailo Yu. A. Memory subsystem performance scaling problems in chip multiprocessors and their solution. Voprosy radio-elektroniki, 2018, no. 2, pp. 23–31. (In Russian).
2. Kozhin A. S. Kontroller pamyati DDR2 SDRAM i ego sistema sinkhronizatsii v sostave sistemy na kristalle Elbrus-S2. Trudy 52 nauchnoi konferentsii MFTI “Sovremennye problemy fundamentalnykh i prikladnykh nauk”. Chast I. Radiotekhnika i kibernetika [Modern Challenges of Fundamental and Applied Sciences. Part I. Radioengineering and Cybernetics. Proceedings of the 52nd Scientific Conference of MIPT]. Vol. 1. Moscow, MFTI, 2009, pp. 55–58. (In Russian).
3. Kim A. K., Perekatov V. I., Feldman V. M. On the way to russian exasistemes: plans of the Elbrus hardware-software platform developers on creation of an exaflops performance supercomputer. Voprosy radioelektroniki, 2018, no. 2, pp. 6–13. (In Russian).
4. Addendum No. 1 to JESD79–4, 3D Stacked DRAM. JEDEC, 2017. Available at: https://www.jedec.org/standards-documents/docs/jesd79-4-1 (accessed 03.07.2019).
5. Choi J. S. Next Big Thing: DDR43DS. Server Forum, 2014. Available at: https://www.jedec.org/sites/default/files/files/ JS_Choi_Server_Forum_2014(1).pdf (accessed 03.07.2019).
6. Dice P. Quick Boot: A Guide for Embedded Firmware Developers, 2nd ed., Berlin, De Gruyter Press, 2017. 284 p.
7. Fruehe J. AMD EPYC Brings New RAS Capability. Moor Insights & Strategy, 2017. Available at: https://www.amd.com/system/files/2017–06/AMD-EPYC-Brings-New-RAS-Capability.pdf (accessed 03.07.2019).
8. Bilyaletdinov I. E., Ometov A. E., Timin L. S. Optimization of parameters of high-speed channels of the processor in order to increase the fault tolerance of the computer complex. Voprosy radioelektroniki, 2018, no. 2, pp. 87–92. (In Russian).

AUTHOR
Igor A. Petrov, senior engineer, MCST JSC, 24, ulitsa Vavilova, Moscow, 119334, Russia, tel.: +7 (499) 135-33-61, e-mail: petrov_i@mcst.ru.

Submitted 29.04.2019; revised 22.06.2019; published online 21.08.2019.