Computer and Information Science; Vol. 13, No. 3; 2020
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Main Scientific and Technological Problems in the Field of Architectural Solutions for Supercomputers

Andrey Molyakov[1]

[1] Institute of Information Technologies and Cybersecurity, Russian State University for the Humanities, Moscow, Russia

Correspondence: Andrey Molyakov, Institute of Information Technologies and Cybersecurity, Russian State University for the Humanities, Moscow, 117534, Kirovogradskaya street, 25/2, Russia.

Received: June 14, 2020 Accepted: July 17, 2020 Online Published: July 24, 2020

doi:10.5539/cis.v13n3p89 URL: https://doi.org/10.5539/cis.v13n3p89

Abstract

In this paper author describes creation of a domestic accelerator processor capable of replacing NVIDIA GPGPU graphics processors for solving scientific and technical problems and other tasks requiring high performance, but which are characterized by good or medium localization of the processed data. Moreover, this paper illustrates creation of a domestic processor or processors for solving the problems of creating information systems for processing big data, as well as tasks of artificial intelligence (deep learning, graph processing and others). Therefore, these processors are characterized by intensive irregular work with memory (poor and extremely poor localization of data), while requiring high energy efficiency. The author points out the need for a systematic approach, training of young specialists on supporting innovative research, improving and developing technological approaches, as shown in paper (Adamov et al., 2019).

Keywords: graphic processors, signal processing processors, multi-tile processors, quantum cellular automata, superconductor electronics

1. Introduction

The revolutionary development, taking into account the global worldwide processes, is the transition from CMOS technologies to Post-Moore's technologies. Among the identified areas, the first two are the most important:

a) Creation of a domestic accelerator processor capable of replacing NVIDIA GPGPU graphics processors for solving scientific and technical problems and other tasks requiring high performance, but which are characterized by good or medium localization of the processed data;

b) Creation of a domestic processor or processors for solving the problems of creating information systems for processing big data, as well as tasks of artificial intelligence (deep learning, graph processing and others), which are characterized by intensive irregular work with memory (poor and extremely poor localization of data), while requiring high energy efficiency.

The motivation for choosing these areas and the other three is determined not only by the needs of the country, but also by the following two factors. On the one hand, the professional level of micro-electronic divisions of the enterprise in the development of specialized VLSI projects has grown significantly in recent years. For example, this is confirmed by the release of the most productive VLSI in recent years in the form of a hybrid 21-core NM6408MP microprocessor (VLSI 1879ВМ8Я) with the NeuroMatrix architecture, which has a good modernization reserve (Chernikov *et al.*, 2018).

Moreover, development of VLSI with great capabilities requires new developments at the architectural level, the levels of system and application software. Not enough attention was paid to these types of work, even on a national scale, and it was decided to try to correct this situation, at least on an initiative basis. Note that this kind of work on innovative architectures, with the possible exception of, perhaps, the Angara project of a supercomputer of strategic importance, has not been deliberately conducted in the country either. The practical result of the Angara project, which, unfortunately, was stopped more than 10 years ago, turned out to be only the Angara communication network (Sluckin *et al.*, 2007). This network and microprocessors "Elbrus" are today the only two products of domestic electronics.

The new products that appeared in foreign products, especially Chinese ones, confirmed the correctness of the

assessment made a year ago on the prospects of developing problem oriented specialized VLSI. In 2019-2020, it is possible to state the emerging new world challenges and, unfortunately, the little-changing situation on work in the field of microelectronics in our country. Changes in the field of urgent applied problems for high-performance systems (neural computing, graph processing and other tasks of artificial intelligence) have manifested themselves in the world. The development of microelectronic technologies ahead of forecasts also occurred (instead of 2020, the limit of 7 nm was reached in 2018, in 2019 – even 5 nm). Both of these phenomena have contributed to an increase in the activity of creating components of foreign microelectronics that are new in architecture and microarchitecture.

In our country the lack of activity on import substitution, the lack of a general organization of work on supercomputer technologies (this has been unsuccessfully discussed for more than 15 years), as well as increased sanctions have led to a rather difficult situation not only in the field of domestic microelectronics, but also in the field of domestic supercomputers in general, even for cluster-type supercomputers using a foreign electronics. Additionally, since the world has become a noticeable trend in the use of responsible supercomputers in precision weapon control systems, especially in their video processing subsystems, the lag was expected to be reflected in the corresponding domestic developments.

2. Development of GPGPU Type Accelerator for Scientific and Technical Calculations

Such devices are used to solve scientific and technical problems with large computing capacity, as well as due to their high performance, and to solve problems of a different type, from cryptographic and neural computing to graph processing. Although, specialized processors would be better suited for these tasks, but GPGPU is easier and more profitable for a number of applications.

Firstly, the peak performance of such a domestic processor-accelerator is estimated by 2027 to be no less than 15 Teraflops on 64-bit floating-point arithmetic, or no worse than 10 times greater than the microprocessor “Elbrus” will have by this time. At the same time, special attention in its development should be given to rather efficient solution of problems on dynamically changing irregular grids, neural computations and graph processing. Although, for such problems it is recommanded to develop processors with limit indicators in the second of the selected areas. Secondly, the following three development options for such a domestic processor-accelerator, which do not exclude each other.

The first option is to follow the research project of the NVIDIA Echelon GPU (Oreste W., 2014) to create a modernized, more uniform version of the GPGPU in the form of a large system based on 64 thread cores. In this option, the architectural principles of multi-core and mass multi-threading remained. According to this option, the company NICEVT has a reserve in the form of a 64-thread microprocessor J7 of the project “Angara” on the FPGA (Mitrofanov V.V., Sluckin A.I., Ejsymont L.K., 2008; Semenov A.A., Sokolov A.A., Ejsymont L.K., 2009; Mitrofanov V.V., Ejsymont L.K., 2008; Zhabin I. *et al.*, 2013).

The second option is a Chinese alternative GPGPU variant, which is accelerator processors developed at the University of Defense Technology of China (NUDT) based on a DSP processor for processing signals like GPDSP FT-Matrix2000 (Chao Y., 2019), FT-Matrix2000 + and FT-Matrix3000 for the planned supercomputer Tianhe-3. These processors are ideologically comparable with the domestic hybrid scalar-vector processor NM6408MP.

The third option is the multi-tile microprocessor of the architectural direction Green Droid (M. Khazraee *et al.*, 2017), which received this name due to the potentially achievable high energy efficiency. It is proposed that each tile of such a processor have a 64-thread core and a static memory of at least 64 KB, which has interfaces for connecting several complex-functional devices oriented to a certain subject area, for example, for neural computing or solving cryptography problems. The main function of the multi-thread kernel of such a tile is to provide dense data loading for such complex functional devices. This solution is more specialized than the first two, but also promises to be more productive and energy efficient.

3. Development of a Strategic Hybrid Processor for Solving Tasks with Intensive Irregular Work with Memory

The main feature is ensuring tolerance of the developed real performance to delays in the execution of memory access instructions, which can be hundreds or thousands of processor cycles for tasks in which such instructions make up the majority, and their memory access addresses are quite unpredictable, vary widely. For these reasons, in conventional processors, data caches of any size become useless for such tasks; page descriptor caches are useless; and significant problems arise due to frequent changes of virtual memory pages in physical memory. Such tolerance in processors of this type is provided by two architectural techniques implemented in them.

The first trick is the basic architectural principle used in such processors mass multithreading, much larger than

in GPGPU. At the same time, the processor supports tracking a large number of memory accesses, thousands in comparison with dozens in conventional processors. This allows such processors to efficiently work with memories with high throughput due to the possibility of simultaneous execution of multiple calls.

The second method is the special organization of the subsystem of virtual memory with a complex process of translating a virtual address into physical. This is required to work with huge amounts of memory, consisting of many physical resources, but located in a single address space.

Processors of this type and systems based on them were called DIS-systems (Data Intensive Systems) appeared about four decades ago, and experimental samples and the results of their research became noticeable in the 90s of the last century (Eisymont L., 1990).

Industrial prototype from Cray (128-thread microprocessor "Threadstorm") appeared in the first decade of the 2000s and turned out to be quite popular among users, especially specialists from the Northwest Pacific National Laboratory of the USA (PNNL) and the Institute of Georgia. The development options for such processors were also studied (Villa O. *et al.*, 2012; Timeo S. *et al.*, 2012). In our country a variant of such a processor was developed in project "Angara".

The pinnacle of this decade in the United States, according to information from the expert community, was the creation of the MRF (Multiprogram Research Facility) unit of the Oak Ridge National Laboratory (ORNL) as part of the closed part of the DARPA HPCS hybrid supercomputer Cray XT5h. In addition to the vector supercomputer segment in the form of Cray X2, it includes a segment with computational nodes with FPGA, as well as a Cray XMT-4 segment with mass-multithread microprocessors Cray Scorpio, which are the development of the Threadstorm microprocessor (it is only known that operations on short and long vectors). This supercomputer was installed at the NSA intelligence center in Utah (Bamford J., 2012; Lakhotia K. *et al.*, 2019). Moreover, it is known from the expert community that similar developments were adopted in China and Japan for similar centers. Actively developing work on graph processing processors based on 3D assemblies of memory modules with built-in layers of logic processing crystals – HMC modules became noticeable. Not only new architectures are being developed, but also new graph partitioning algorithms to obtain greater data localization when they are processed in highly parallel systems.

4. Conclusion

The selected problems are considered from different points – analytical, educational and individual research activities among young professionals and partner organizations in industry. We believe that these areas and their motivation should be considered by experts of the relevant commissions and councils in industry. The main conclusions are as follows:

a) The new challenges of 2019-2020 demanded to expand the list of work on import substitution. In this regard, the main directions of innovative research were formulated. Some components of supercomputers have become in demand in the control systems of promising domestic combat drones;

b) The primary task is the development of an analogue and / or an alternative version of GPGPU. Three solution options were identified: universal homogeneous mass-multithread (like GPGPU Echelon); scalar vector GPDSP; multi-file with multi-thread cores in tiles and complex functional blocks connected to them (Green Droid);

c) Formed two important fields of researching. The first, universally recognized in relevance, is the creation of systems with deep learning (neuroalgorithms). The second, rapidly gaining relevance and in the future, which will obviously be fundamental, is the creation of scalable processing systems for large graphs.

d) Immediately, without waiting for decisions on the part of government agencies, to begin the process of training young specialists in the form of interdepartmental seminars and etc.;

e) Particular attention should be paid to information and analytical work in selected areas;

f) New research projects require study and analysis, although so far with a lower priority: communication networks of a new topology with a small diameter (intracrystal and interstitial), next generation architecture, microarchitecture and circuitry;

g) A common phenomenon for these studies was that among specialists: products based on superconducting electronics would require a departure from conventional circuitry and microarchitecture towards cellular automata;

h) In architectural terms, it can be considered as the next level of generalization of creating graph processing processors, since the mathematical basis of the instruction system of such processors is no

longer operations on sparse matrices, but a whole functional language for processing graphs with fragments in the form of tree-like symbols free strings and basically without the side effects of performing functions.

Acknowledgments

Identify grants or other financial support (and the source, if appropriate) for your study; do not precede grant numbers by No. or #. Next, acknowledge colleagues who assisted in conducting the study or critiquing the manuscript. Do not acknowledge the persons routinely involved in the review and acceptance of manuscripts 羊 peer reviewers or editors, associate editors, and consulting editors of the journal in which the article is to appear. In this paragraph, also explain any special agreements concerning authorship, such as if authors contributed equally to the study. End this paragraph with thanks for personal assistance, such as in manuscript preparation.

References

Adamov, A. et al. (2019). Main problem directions in the field of domestic element base of supercomputers. *Cybersecurity Issues*, *4*, 2-12. https://doi.org/10.21681/2311-3456-2019-4-02-12

Adamov, A. A., Pavlukhin, P. V., Bikonov, D. V., Eisymont, A. L., & Eisymont, L. K. (2019). Prospective general purpose and specialized accelerator processors alternative to modern GPGPU. *Cybersecurity Issues, 4,* 13-21. https://doi.org/10.21681/2311-3456-2019-4-13-21

Bamford, J. (2012). *The NSA is building the country's biggest spy center (watch what you say).* Retrieved from https://www.wired.com/2012/03/ff-nsadatacenter/

Chao, Y. et al. (April 2, 2019). A Novel DSP Architecture for Scientific Computing and Deep Learning. *IEEE Access*, *7*, 36413-36425. https://doi.org/10.1109/ACCESS.2019.2905302

Chernikov, V. M., & Viksne, P. E. P. (2018).Improving the characteristics and expanding the areas of application of transteraflops VLSI of the NeuroMatrix family. *High availability systems*, *14*, 28-34.

Eisymont, L. K. (1990). Computers for processing symbolic information. *Foreign radio electronics*, *4*, 3-28.

Khazraee, M., Gutierrez, L. V., Magaki, I, & Tailor, M. B. (May/June, 2017). Specializing a Planet's Computation: ASIC Clouds. *IEEE Micro*, 62-69. https://doi.org/10.1109/MM.2017.49

Lakhotia, K. et al. (2019). GPOP: A scalable cache- and memory- efficient framework for Graph Processing Over Partitions. https://doi.org/10.1145/3293883.3299108

Mitrofanov, V. V., & Eisymont, L. K. (2008) Element base and architecture of high-performance multiprocessor computing systems, promising strategic and embedded supercomputers. *Dynamics of radio electronics, 2nd issue. Ed. Borisova Yu.I.* ISBN: 978-5-94836-195-6, 70-76.

Mitrofanov, V. V., Slutskin A. I., & Eisymont, L. K. (2008). Supercomputer technologies for strategically important tasks. *Electronics: NTB*, *7*, 66-79.

Oreste, W. et al. (November 16-21, 2014). Scaling the Power Wall: A Path to Exascale.*Supercomputing Conference (SC14)*. Retrieved from https://www.cs.utexas.edu/users/skeckler/pubs/SC_2014_Exascale.pdf

Semenov, A. A., Sokolov, A. A., & Eisymont, L. K. (2009). Globally Addressable Memory Architecture of a Multithread Streaming Supercomputer. *Electronics: NTB*, *1*, 50-56.

Slutskin A.I., Eisymont L.K (2007). Russian supercomputer with globally addressable memory. *Open systems Journal*, 9, 42-51. Retrieved from http://www.osp.ru/os/2007/09/4569294/

Timeo, S. et al. (August 2012). Designing Next-Generation Massively Multithreaded Architectures for Irregular Applications. *COMPUTER*, *45*, 53-61. https://doi.org/10.1109/MC.2012.193

Villa, O. et al. (13 Feb 2012). Fast and Accurate Simulation of the Cray XMT Multithreaded Supercompute. *IEEE Transactions on Parallel and Distributed Systems*, *9*, 2266-2279. https://doi.org/10.1109/TPDS.2012.70

Zhabin, I., Makagon, D., Simonov, A., Syromyatnikov, E., Frolov, A., & Shcherbak, A. (2013). Crystal for "Angara". *Supercomputers*, *4*, 46-49.