UDC 004.272.26, 004.274
DOI: 10.15587/1729-4061.2016.71460

THE DEVELOPMENT OF MEANS OF DEFINITION OF THE OPTIMUM RATIO OF COMPUTATIONAL ALGORITHM AND THE RECONFIGURABLE STRUCTURE

I. Klymenko
PhD, Associate Professor*
E-mail: iklymenko@yandex.ua
O. Holovko*
M. Hilliaka*
E-mail: maxymhammer@yandex.ru
Y. Mytsio*
E-mail: saturn4er@gmail.com
*Department of Computer Engineering
National Technical University of Ukraine
“Kyiv Polytechnic Institute”
Peremohy ave., 37, Kyiv, Ukraine, 03056

Запропоновано спосіб визначення оптимального співвідношення обчислювального алгоритму і структури обчислювального середовища за критеріями часових та апаратних обмежень реконфігуровних обчислювальних систем, з врахуванням комунікаційних затримок фізичного рівня кристалів програмованих логічних інтегральних схем (ПЛІС). Запропонована нова стратегія взаємної адаптації обчислювальних алгоритмів і обчислювального середовища, що дозволила підвищити ефективність реконфігуровних обчислювальних систем для рішення задач надвеликої розмірності

Ключові слова: реконфігуровні обчислення, зернистість обчислень, програмовані логічні інтегральні схеми, комунікаційні затримки

Предложен способ определения оптимального соотношения вычислительного алгоритма и структуры вычислительной среды по критериям временных и аппаратных ограничений реконфигурируемых вычислительных систем, с учетом коммуникационных задержек физического уровня кристаллов программируемых логичных интегральных схем (ПЛИС). Предложена новая стратегия взаимной адаптации вычислительных алгоритмов и вычислительной среды, которая позволила повысить эффективность реконфигурируемых вычислительных систем при решении задач сверхбольшой размерности

Ключевые слова: реконфигурируемые вычисления, зернистость вычислений, программируемые логические интегральные схемы, коммуникационные задержки

1. Introduction

In recent years, the development of the field-programmable gate arrays (FPGA), which are dynamically reconfigurable, made preconditions and new possibilities for increasing the efficiency of parallel computing systems by making it possible to reconfigure a computing system at run-time [1–3]. However, an effective implementation of tasks mapping on the dynamically reconfigurable system is connected with great inefficient wastes of time and performance during its reconfiguration [1, 2]. The problem is about the hardware resource limitations of the FPGA, overcoming of which, especially during the solution of tasks of a big size, additionally increase reconfiguration overhead. Features of physical processes of signal spreading at the chip level of the FPGA have a great negative impact on mapping efficiency, at the same time any of the known overhead reduction methods include these figures. In effect, this can negate any hardware accelerations.

In summary, the relevance of this topic is caused by the need of the development of new or of improving the known tools of tasks mapping on the computational structure of dynamically reconfigured computational systems, which include hardware, constructive and technological limitations of the FPGA chips and can be effectively used to solve tasks of big size.

2. Analysis of literature and the given problems

There are a lot of methods and technologies for reconfiguration overhead reduction. The most known are resource reuse [1], caching of configuration data [2], forward reconfiguration [4], hardware tools of input/output acceleration [3] and optimization of the virtual structure of configuration data [5]. Each of them is based on the maximum possible reconfiguration acceleration of type “Best Effort” without any optimization of space solution and excluding hardware and technological limitations of the FPGA. Overcoming the space limitations of the FPGA is done with standard tools, for example, defragmentation of the computational surface of the FPGA [1, 2], loading out non-critical configuration [1, 2, 4], which brings an additional overhead to the reconfiguration.

Granularity is considered as an effective space solution in the field of parallel computation [6]. The reconfiguration possibilities of the computing structure give perspectives to build an ideal computing structure for each task by varying the granularity level. Well-known methods and tools to vary the granularity level, developed for fixed computing structures [6], do not consider features of reconfigurable computations and cannot be effective for such a usage.

In the field of reconfigurable computing systems, the problems of varying the granularity level are described in the following topics:

[7] – tasks mapping is done based on the idea of defining the needed amount of computational resources from the global reconfigurable computing space, which is shared among cores, which consists of predefined sets of fine-grained and coarse-grained modules**;**

[8, 9] – the principle of reconfiguration based on the usage of hardware instruction set extensions (ISEs) to accelerate functional core;

[10] – the principle of reconfiguration, based on the usage of coarse-grained reconfigurable architectures. These known solutions of tasks mapping are based on defining the minimum amount of hardware to provide the required solution time for the algorithm by reconfiguring connection channels between issued computing structures. The granularity variation is limited with adjustment of the computational algorithm to the defined form, caused by the system structure. The problem of adopting a computing structure to the requirements of the task, according to the paradigm on reconfigured computations, is not discussed in the overviewed topics. Also, the problem of hardware limitations of the computing structure, which appears in case of incompatible increment of task size, is neither discussed**.**

Despite the fact that extreme level of modern FPGAs integration allows to abstract from space limitations of the FPGA chips, space, technological and physical characteristics of the FPGA chips have a certain impact on the efficiency of reconfigurable computations during the task solution of big size. The computational "seed" size is affected not only by traditional parameters of parallel algorithm complexity and FPGA space parameters, but is also greatly affected by transfer delays of the inner channels of data transmission. These issues have never been studied before.

It should be noted, that known tools of task mapping on the parallel computing structure are developed for the fixed architectures or switched computing environment, cannot be effectively used to solve tasks of big size on reconfigurable computing systems, which have certain software or hardware limitations. It proves unsolved issues in the field of reconfigured computing systems and also the need and expediency of the researches done in this work.

3. Goal and tasks of the research

The purpose of the research is the improvement of efficiency of dynamically reconfigured system by the new tasks mapping strategy, which is based on the mutual adaptation of computing algorithm and computing environment.

To achieve the study goal, it is needed to solve the following tasks:

– determining and researching criteria of fast reconfigurable computing FPGA space;

– development of the way to determine an optimum ratio of the structure of reconfigurable computing FPGA space and computing algorithm based on the determined performance criteria;

– development of the library of software cores, which allows to effectively vary the computational granularity;

4. Materials and methods of the research and defining an optimum ratio between the structure of a computing environment and a computing algorithm

4. 1. The research objects, programmable tools and equipment, which are used during the research

Functional blocks of hardware tasks, on the base of which the modeling and researches of a new strategy of reconfigurable computations were done, are synthesized with Verilog – the hardware description language, and implemented on various families of FPGA Cyclone of Altera Company. To analyze time characteristics and test the correctness of functional blocks, the Altera Development Kit DE2 was used.

4. 2. Reasoning of an optimum ratio between a computing algorithm and the structure of reconfigurable computing environment

Acceleration index of reconfigurable computations:

$$\rho=\frac{T_{SW}}{T_{Rconf}+T_{HW}}, \tag{1}$$

where T_{SW} – time of task computation with one core, T_{HW} – time of task computation with hardware tools, T_{Rconf} – reconfiguration time of computing structure, defined by the authors in the previous topics [4, 11]. According to this, one of the efficiency criteria of reconfigurable computing systems is the speed of computational task solution with the hardware tools T_{HW}. Besides this, the structure of the computing system environment has a significant impact on the calculation time. Calculation acceleration is traditionally achieved by increasing the calculation "seed", which leads to the minimization of the calculation complexity of parallel processed functions and reduction of the transferred data. But such a reconfiguration degree, which corresponds to an ideal client computation efficiency, leads to the great overhead to rebuild the computing structure on the FPGA surface.

In this topic, it is proposed to solve the problem of reconfiguration optimization by defining an optimum ratio between a computation algorithm and the structure of the computing environment, which leads to the minimization of the efficiency parameter T_{HW} (1). As the optimization criteria, there is the time limitation, which is common for reconfigurable calculations $((T_{Rconf}+T_{HW})<T_{SW})$, and hardware resource limitations of the FPGA. Definition of an optimum ratio is proposed in a way, which is based on the notion that to achieve the maximum performance (with $T_{HW}\downarrow$), the granularity of the computation algorithm, where S – the granularity coefficient of the algorithm, must be equal to the granularity of the computing space, where K – the granularity coefficient of the computing space (Fig. 1).

In Fig. 1, there is a dependency between the computation time and the granularity degree of the computing space and definition of the optimal time parameter, which corresponds to the notion

$$S=K. \tag{2}$$

The granularity coefficient of an algorithm corresponds to the ratio of computation complexity of the algorithm to the data transmission size:

$$S=\frac{V_{Count}}{V_{IO}},$$

where V_{Count} – time complexity of the computation algorithm, what equals to the number of solved operation, V_{IO} – number of performed data transmission operations.

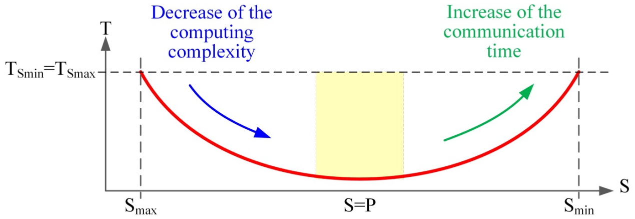

Fig. 1. Area of an optimum ratio between a computing algorithm and a structure of a computing space

Then for some algorithm, let's define the minimum granularity coefficient, where the algorithm has the maximum parallelism, and the maximum granularity coefficient and when the functional unit performs all the calculations:

$$S_{min}=\frac{1}{n},\ \ S_{max}=\frac{n}{1}.$$

Let's give a notion of k – algorithm granularity of the computation task, which has the size of n and $(k \le n)$. Then, the granularity coefficient of the algorithm equals to

$$S_{task}=\frac{n/k}{k}=\frac{n}{k^2}. \tag{3}$$

Ratio optimization of the computing algorithm and the computation structure has the goal to increase the performance of parallel computations. To specify optimal granularity, let's define a performance index of a reconfigurable computing space for the FPGA, which describe a ratio between processing speed (V_{Count}) and transmission speed (V_{IO}):

$$K_{FPGA}=\frac{V_{Count}}{V_{IO}}. \tag{4}$$

4. 4. Reasoning of speed criteria of the reconfigurable computing space of the FPGA

To estimate the performance of a functional unit, which is a hardware implementation for some task, let's apply a common parameter of job frequency (F_{func}), which is defined as the number of operation per time unit

$$F_{func}=\frac{1}{t_{func}},$$

where t_{func} – the time to solve the computation task. For estimation of the data processing performance of the whole reconfigurable structure (F_{FPGA}), we get the following statement, where k\` – the number of function units:

$$F_{FPGA}=F_{func}\times k`. \tag{5}$$

The data transmission speed is limited by technical parameters of the input/output interface of the FPGA and data transmission channels between the memory and the reconfigurable structure and does not overflow the value of F_{IO} – the work frequency of the external memory. Then, the load frequency of functional units of hardware tasks F_{IO_func}, implemented on the FPGA surface, is defined as:

$$F_{IO\_func}=\frac{F_{IO}}{m}. \tag{6}$$

Based on the expressions (4)–(6), we get the performance index of the reconfigurable FPGA space:

$$K_{FPGA}=\frac{F_{FPGA}}{F_{IO\_func}}=\frac{F_{func}\times k`}{F_{IO}/m},\ \ K_{FPGA}\ge 1. \tag{7}$$

4. 5. Way to find an optimum ratio of a computing algorithm and the structure of a reconfigurable computing space

Based on the expressions (2), (3) and (7), we can get the following expression to find the granularity, which shows an optimum ratio between a computation algorithm and the structure of reconfigurable computing space from the sight of the time minimization of reconfigurable computation including the space limitations of the FGPA chip (an optimum granularity):

$$k=\frac{n}{F_{func}}\times\sqrt{\frac{F_{IO}}{F_{func}}}, \tag{8}$$

where n – the task size and m=n/k is true for regular computing structures. This is a special case, which covers regular computing structures, particularly to compute functions of linear algebra, matrix equations and so on.

4. 6. The physical nature of delays, which affects the performance of reconfigurable computing space

Let's make an assumption, that physical properties of the FPGA chips cause the influence of the size of functional units on time delays during the signal transmission with the inner connection channels. Let's define the delays of inner reconfigurable structure, which are described as:

$$d_{FPGA}=d_{comm}+d_{io},$$

where d_{comm} – inner communication delays, d_{io} – duration of the delay to transmit the data array from memory to the FPGA.

Hardware computation implementation allows to effectively implement the chunked data transmission. Then, theoretically, the time to perform some algorithm with the minimum (S_{min}) and maximum (S_{max}) granularity degree may be reduced by decreasing the number of memory access delays:

$$T_{S_{min}}=[n\times[m_{min}\times(t_{io}+t_0)+\tau]],$$

$$T_{S_{max}}=[n\times\tau+[t_0+m_{max}\times t_{io}]], \tag{9}$$

where τ clocks – time to process one operation, m – number of transmitted machine words while processing one “grain” of an algorithm, t_{io} – time to transmit one word of data, t_0 – time of a memory access. The square brackets indicated to an atomic operation, what means to the transmission of one data block. From the equation (9), it can be inferred that the processing time decreases with the incensement of the algorithm granularity ($T_{S_{max}}<T_{S_{min}}$). Given the fact that the

number of clocks needed to write or read data from memory is proportional to the number of elements in a processed matrix, we need to find the real dependency of delays from configuration the reconfigurable computing space.

Reconfigurable computations at the physical level of the FPGA chip is controlled by the state machine synthesized from the code written in Verilog. Thus, the nature of inner communication delays substantiates the execution of the algorithm what is accompanied by transmission of the corresponding control signals among inner communication channels.

5. An experimental definition and finding the value of the speed criteria of the reconfigurable computing space

5. 1. Research of the delays, caused by the physical properties of reconfigurable computing space

We performed experiments with calculations of matrix functions of different size, which are characterized by different data block sizes, which are loaded for calculations. As the result, we have got the following dependency of data transmission delays from the size of the functional block for different external data sizes (Fig. 2).

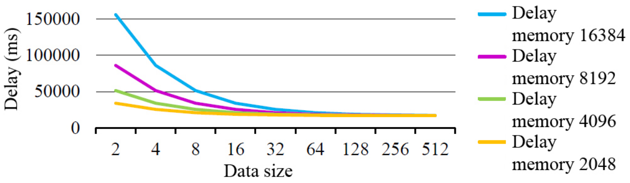

Fig. 2. Analysis of delays for chunked data transmission

Based on the results, it was figured out that the block size from 2 to 32 bytes is characterized by long delays, which decrease with the increasing data block size. It can be explained by the following expression, inferred from the theoretical statement (9): $d_{io} = k \times t_0 + m \times t_{io}$, where k – the number of data blocks, t_0 – the time to set the memory pointer to the start position and the time to receive the signal about the ready signal; the read/write time corresponds to one clock $(t_{io} = 1\tau)$. The following incensement of the data block size does not reduce the delay and actually the delay stops to depend on the data block size, because the value of m becomes so big that the multiplication $(m \times t_{io})$ plays a bigger role, than $k \times t_0$.

We have researched the impact of space parameters of functional blocks on time delays during the signal transmission among inner communication channels. In Fig. 3, there are the results of the research about the frequency of functional block for implementation of the computing algorithm of linear algebra. On the diagrams, it can be seen that computation performance decreased by 50 % while increasing the size of the computing function and accordingly while increasing the space parameters of the functional block. At the same time, the deceleration intensity depends also on technical characteristics of an integrated circuit.

The results of the research about the efficiency of changing the granularity according to the proposed expression (8) are shown in Fig. 4.

Researches are done for two families of the FPGA chips with high performance characteristics, which have a reasonable price and are actual for use within labs.

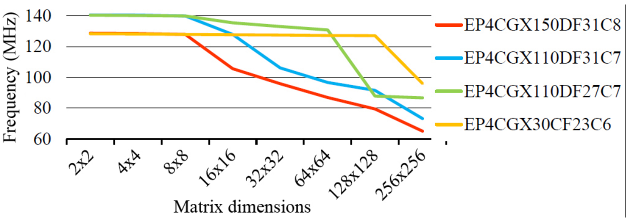

Fig. 3. Experimental performance estimation of functional blocks

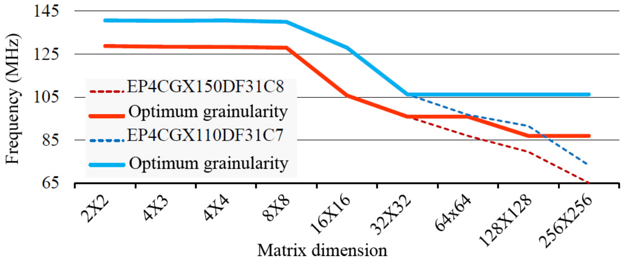

Fig. 4. Research of the change of computation granularity

5. 3. Modeling the way to find an optimum ratio between a task algorithm and a structure of reconfigurable computing space

Within this topic, we have made researches for the series of computing algorithms, shown as the macro dataflow graphs, the structure of which is generated randomly [4]. In the roots of the examined graphs of algorithms, macro tasks are located, which actually are the random set of matrix operations of different size. Based on the developed library of functional blocks, we have made an appropriate modeling of the proposed way to find an optimum ratio of a computing algorithm to the structure of a computing space based on the expression (8). Based on our researches, we have got the dependency of time of reconfigurable computations on the size of the similar tasks. In Fig. 5, there is an effect of usage of proposed methods [4] to speed up reconfiguration with incorporating the proposed methods.

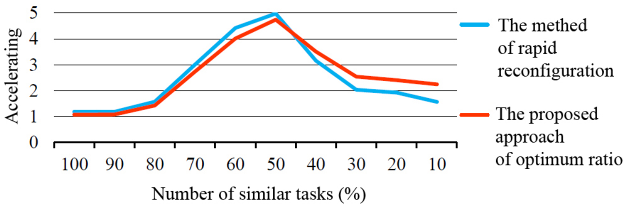

Fig. 5. Research of a computation acceleration rate

At critical areas, which correspond to the occurrence of space limitations of the FPGA chip, there is a sharp reduction of intensity of reconfiguration by 85 % [4]. The proposed method of finding an optimum ratio of a computing algorithm to computing space reduce the influence intensity of space limitation on the computation performance in particular by 10 % on average.

6. Discussing the research and experiments results

The researches done in the topic approve the theoretical notion about an impact of physical processes features at the FPGA chips on the efficiency of reconfigurable computations. Graphics I and II at the diagram (Fig. 2) show that in case of implementation of the maximum granularity coefficient while increasing the task size there is a great decrease of performance. Graphics III and IV show the results of applying the proposed method to find an optimum ratio of computing algorithm to the structure of computing environment. A granularity optimization of reconfigurable computations based on the proposed method leads to the increase of computation from 9 % to 15 % on average and depends on technical properties of the FPGA chips.

The research of the efficiency of blocked data transmission, which theoretically leads to the linear decrease of input/output delays due to the increase of the data block size, proves that the increase of the data block size for matrix operation of the size more than 128 can be neglected for the input/output delay.

The library of functional blocks for matrix operations was developed and examined, particularly we worked with the operation of matrix addition and multiplying a matrix by a scalar. To get the most objective experimental evaluation of values of delays of the FPGA computing space, operations, which are evaluated for any tasks size in one tact, were researched. But such values are true for any computing operation because the researched delays appear at data input/output and computation control stages.

The researches done in this topic follow the researches, which are described in details in previous topics [4, 5] and aimed at increasing the reconfigurable computations efficiency by space solution optimizations. As the further researches, we plan to expand the library of the functional block to solve the wide range of tasks of big and super big size on reconfigurable computing units of the FPGA.

7. Conclusions

Performance criteria of reconfigurable computing space, which are researched in this topic, are based on finding an optimum ratio between a computing algorithm and the structure of the computing environment in terms of minimization of reconfigurable time, minimization of data transmission delays and space and structural and functional limitations of reconfigurable computing units based on the FPGA. Consideration of the defined criteria during solving the task of mapping algorithms on reconfigurable computing structure allows to take into account critical parameters such as signal spread delays on the physical level of the FPGA chips, the negative impact of which, based on the research, linearly depends on the “seed” of computations.

The proposed method to find an optimum ratio between a computing algorithm and the structure of reconfigurable computing space, which is based on the defined performance criteria, allowed implementing a new strategy of task mapping, based on the mutual adaptation of the computing algorithm and the computing environment. It allows to increase the client efficiency of reconfigurable computing systems built on the FPGA basis by 10 % based on the results of the research.

The library of functional cores to solve tasks of linear algebra and matrix operations, proposed and build on the FPGA, allowed providing the set of functional blocks with optimum characteristics according to defined performance criteria of reconfigurable computing space, which provides an efficient way to vary the computation granularity.

References

1. Bassiri, M. M. Mitigating Reconfiguration Overhead In On-Line Task Scheduling For Reconfigurable Computing Systems [Text] / M. M. Bassiri, H. S. Shahhoseini // 2010 2nd International Conference on Computer Engineering and Technology. – 2010. – Vol. 4. – P. 397–402. doi: 10.1109/iccet.2010.5485509
2. Al-Wattar, A. Efficient On-line Hardware/Software Task Scheduling for Dynamic Run-time Reconfigurable Systems [Text] / A. Al-Wattar, S. Areibi, F. Saffih // 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum, 2012. – P. 401–406. doi: 10.1109/ipdpsw.2012.50
3. Liu, S. Achieving Energy Efficiency through Runtime Partial Reconfiguration on Reconfigurable Systems [Text] / S. Liu, R. N. Pittman, A. Forin, J.-L. Gaudiot // ACM Transactions on Embedded Computing Systems. – 2013. – Vol. 12, Issue 3. – P. 1–21. doi: 10.1145/2442116.2442122
4. Kulakov, Y. O. Devising statistic models of milking duration on the conveyor milking machines [Text] / Y. O. Kulakov, I. A. Klymenko, M. V. Rudnytskyi // Eastern-European Journal of Enterprise Technologies. – 2015. – Vol. 4, Issue 4 (76). – P. 25–29. doi: 10.15587/1729-4061.2014.28951
5. Kulakov, Y. O. The multilevel memory in the reconfigurable computing system [Text] / Y. O. Kulakov, I. A. Klymenko // Visnyk NTUU «KPI». Informatyka, upravlinnia ta obchislyuvalna technika. – 2014. – Vol. 61. – P. 18–26.
6. Levchenko, R. I. A System of Automatic Dynamic Paralleling of Computations for Multiprocessor Computer Systems with Weak Connection (DDCI) [Text] / R. I. Levchenko, O. O. Sudakov, S. D. Pogorelij, Y. V. Bojko // USiM. – 2008. – Vol. 3. – P. 66–72.
7. Ahmed, W. Adaptive Resource Management for Simultaneous Multitasking in Mixed-Grained Reconfigurable Multi-core Processors [Text] / W. Ahmed, M. Shafique, L. Bauer, J. Henkel // Proceedings of the 9th International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2011. – P. 365–374.
8. Koenig, R. Kahrisma: A Novel Hypermorphic Reconfigurable-Instruction-Set Multi-grained-Array Architecture [Text] / R. Koenig, L. Bauer, T. Stripf, M. Shafique, W. Ahmed, J. Becker, J. Henkel // 2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010), 2010. – P. 819–824. doi: 10.1109/date.2010.5456939
9. Sourdis I. Resilient Chip Multiprocessors with Mixed-Grained Reconfigurability [Text] / I. Sourdis, D. A. Khan, A. Malek et. al. // IEEE Micro. – 2016. – Vol. 36, Issue 1. – P. 35–45. doi: 10.1109/mm.2015.7
10. Yin, S. Memory-Aware Loop Mapping on Coarse-Grained Reconfigurable Architectures [Text] / S. Yin, X. Yao, D. Liu et. al. // IEEE Transactions on Very Large Scale Integration (VLSI) Systems. – 2016. – Vol. 24. – P. 1895–1908.
11. Klymenko, I. A. The method of optimization reconfiguration for the dynamic reconfigurable computer [Text] / I. A. Klymenko // Visnyk NTUU «KPI». Informatyka, upravlinnia ta obchislyuvalna technika. – 2015. – Vol. 63. – P. 93–100.