Architectures of high-performance VLSI for custom computing systems

I E Tarasov

MIREA - Russian Technological University, 78, Vernadsky ave., Moscow, 119454, Russia

E-mail: tarasov_i@mirea.ru

Abstract. The article deals with the design of specialized VLSI designed for the construction of high-performance computing systems. Common options for building such systems are general-purpose processors (CPUs), graphics accelerators (GPUs) that work in the mode of general-purpose computing acceleration, and FPGAs that are configured to implement a specific task or a narrow class of tasks. If there is a demanded class of tasks with specific requirements for computing resources, it is possible to develop a specialized VLSI that is more effective than CPU and GPU in terms of the use of crystal resources, and more productive than FPGA due to the use of hardware-implemented connections instead of programmable cells and switched trace lines. Taking into account modern trends in microelectronics and existing technological lag from leading global manufacturers, an important role is played by the choice of VLSI architecture as well as the use of architectural and circuit design solutions that reduce technical risks and provide solutions to urgent problems for Russian companies.

1. Introduction

Currently used computing systems are based on three main components: general-purpose x86 processors, graphics accelerators (GPUs), and programmable logic chips with FPGA architecture. Along with these subclasses, it is possible to consider specialized VLSI with a configurable or programmable structure that increases the effectiveness of the armed forces in the selected subclass of the task. Such VLSIs in general are similar to the idea of a graphics accelerator (GPU), the main purpose of which is to solve coordinate transformation problems for three-dimensional graphics, which badly corresponds to the architecture and capabilities of general-purpose processors. By analogy with a GPU, a similar VLSI with a mass-parallel architecture can be developed, having significant differences from the GPU in the implementation of computational nodes, which will ensure the best efficiency of such VLSI in the corresponding classes of tasks.

Figure 1 shows the main classes of computational architectures designed to build high-performance computing systems. While for general purpose processors (CPUs) the main feature is high performance of computations in several computational threads, for FPGA and vector processors, the main effect of increasing productivity is based on using a large number of relatively simple subsystems.

This paper presents an overview of perspective architectures for VLSI high-performance computing systems in applications which requires custom hardware for highly-specialized classes of algorithms.
2. Problem overview

Announced in 2018, Xilinx’s Versal [1] compute platform combines the classes shown in Figure 1, providing the ARM Cortex-A processor system as a CPU, the standard matrix of FPGA programmable cells, and the subsystem for FPGA chips for the first time vector processors. This combination of architectures on a single chip allows flexible distribution of the computing load between the components of the computing system, choosing the hardware platform that is optimal for each type of task. A similar heterogeneous architecture can also be applied to newly developed VLSI, ensuring compliance with several classes of tasks already at the stage of choosing the VLSI architecture and the computing system as a whole.

Considering the target classes of tasks, it can be noted that obtaining a solution competing with modern GPUs is possible if the data formats differ significantly from those used in the GPU processor elements, as well as reducing the amount of data stored in the VLSI. Such conditions correspond, for example, to the tasks of digital signal processing, including a program-dependent radio, streaming image processing, and neural networks of deep learning. In these cases, processing of integer input data (8–16 bits) is used, and some of the processing algorithms used do not require storing large amounts of data.

3. Designing VLSI for High Performance Computing Systems

The VLSI design process has a number of features, defined by objective technical and economic constraints.

The following trends can be distinguished in the world microelectronics:

1. The continuing decrease in technological standards with the release of VLSI saves the effect of increasing the density of components on a chip, but this is not accompanied by a proportional increase in the clock frequency.

2. An increase in the density of components leads to an increase in the heat output, which cannot be removed from the crystal due to the heat-conducting characteristics of the plate itself (regardless of the cooling system used). This limits the power consumption of 0.5–0.7 W/mm². At the same time, methods of reducing power consumption were used for the technological node of 90 nm, which led to the emergence in 2012 of the concept of ‘dark silicon’ [2], which implies that to maintain performance, only a small part of the crystal can perform useful actions on data processing.

3. The increase in the area of the VLSI limits the size of the region, for which the qualitative placement of the clock network is possible. This leads to a ‘globally asynchronous, locally synchronous’ (Globally Asynchronous, Locally Synchronous, GALS) architecture. For example, starting with the 20 nm technology node (the UltraScale family), Xilinx uses independent clock regions that cannot physically use the global clock network [3].

4. With decreasing technological standards, the influence of the variability of the parameters of sold components increases. Thus, the process with the standards of 28 nm of TSMC implies the presence of
at least 5 groups of technological libraries, which differ in ratios of speed, area and power consumption: HP, HPL, HPC, HPM, LP. As part of the CAD process of microelectronics, automatic selection of components that meet the requirements of the developer for speed is used, but this is due to an increase in power consumption. It should be noted that this is not about raising the so-called. dynamic consumption due to an increase in clock frequency, and an increase in the coefficient of proportionality, which may be non-linear in nature. In combination with the existing restrictions on the power dissipated by the crystal, this forms a set of requirements for the architecture and technical implementation of VLSI nodes, which preclude a sharp increase in power consumption due to the allowed architectural or circuit design errors.

The choice of a limited subclass of tasks is the basis for obtaining VLSI using hardware resources more efficiently than mass-produced computing devices. Taking into account the fact that technology centers capable of designing the VLSI topology are limited in Russia, and the development of the microelectronic element base could be required by a wide range of organizations, the VLSI design at the system level, which is performed without the involvement of specialized microelectronics organizations, becomes especially relevant. and without substantial financial investments.

One of the options for the practical implementation of the design methodology for problem-oriented VLSI is given in [4]. The technique is based on the following provisions:

- high-level modeling tools are used (including high-level programming languages);
- in the process of architecture design, ready-made models of the system architecture are used, freeing developers from the need to independently develop such models;
- a library of components (large blocks) of problem-oriented VLSI with known characteristics is used;
- when developing new components, recommendations on the implementation of models are used, including: a list of allowable model transactions that can be implemented schematically in the selected hardware basis, as well as recommendations on priority use of transactions based on the analysis of component libraries and evaluating the area, power consumption and propagation delay of signals.

The use of high-level models significantly increases the speed of modeling [5], although it requires ensuring the adequacy of the model used for future VLSI.

An example of a transaction that is valid at the level of the program model, but not implemented at the level of register transfers (RTL), may be the operation of division or calculation of the transcendental function. Despite the fact that these operations are implemented in most programming languages, they do not correspond to hardware structures that implement computations per clock cycle based on modern circuit design approaches. Such operations need to be emulated by cycle-accurate approach, or replaced with equivalent models imitating the result obtained through the corresponding number of cycles. It is also obvious that at the level of the system model it is necessary to exclude operations that form a conflict when accessing resources, which is monitored quite simply - by monitoring no more than one write operation to each of the modeled resources.

At the topological level, a number of process-dependent factors should be taken into account, determined individually for the project and the type of technological process selected. These may include not entirely obvious solutions, determined by the ratios of valve delays, delays on tracing lines, as well as the dependence of the input power bridges on the set delays on individual valves. The last factor has an unobvious, but important influence in the process of topological implementation of VLSI (but not projects based on FPGA), as in the process of optimizing the CAD project, VLSI performs replacement of individual valves in order to provide the required overall delay characteristics. It is characteristic that even such a node as an adder can cause an unpredictable increase in power consumption if its delay is set at less than a certain limit. This issue is widely covered in modern publications devoted to the optimization of topological representations of individual digital nodes (for example, [6 - 10]).

The influence of the topological level at the present time can no longer be ignored in the development
of high-performance VLSI with a parallel architecture. While FPGA-based projects are fairly widespread in Russia, the circuitry techniques used for this element base can be unreasonably transferred to the VLSI basis with an over-consumption of energy consumption. The adder module already mentioned in the FPGA uses dedicated hardware resources (fast carry chain), i.e. from the point of view of the schema developer, this is an optimized (“accelerated”) resource compared to nodes running on the basis of LUT. Fixed hardware implementation of accelerated transfer circuits in FPGA causes constant parameters of delay and power consumption, which depends only on the fact of using or not using such a resource. At the same time, when transferring this circuit to VLSI, it is possible to balance the overall signal propagation delay, where the adder module is only one of the circuit elements that form the delay of the critical signal path. As a result, reducing the delay at the adder can cause a nonlinear increase in its energy consumption. If we are talking about the parallel operation of many such nodes, the growth of energy consumption can no longer be neglected. In contrast, single-core processors can allow for an increase in power consumption in the critical circuit, since this does not lead to a proportional increase in the power consumption of the entire chip.

4. Conclusion
Thus, in the process of creating VLSI for high-performance computing systems, you can use a design technique based on the following provisions:

- the choice of the VLSI architecture is made on the basis of system modeling taking into account the peculiarity of the topological implementation of the VLSI - the problem of "dark silicon" and distributed architecture with localized synchronous clock trees;
- specialization is performed for the selected subclass of model problems in order to improve the efficiency of the new architecture in comparison with the competing mass solutions produced by advanced technology;
- co-optimization of software and circuit solutions is applied, including taking into account the features of the selected topological library.

When building high-performance computing systems, the accelerator of computations with a parallel architecture has a great influence on the final characteristics. Thus, the choice of the optimal accelerator architecture for implementation in VLSI is an important design stage, where at the early stages it is necessary to use system-level modeling to test the effectiveness of the algorithms.

References
[1] Versal: The First Adaptive Compute Acceleration Platform (ACAP). WP505 (v1.0) October 2, 2018 Retrieved from: www.xilinx.com.
[2] Esmaeilzadeh H, Blem E, Amant R S., Sankaralingam K, Burger D 2011 Dark Silicon and the End of Multicore Scaling. Appears in the Proceedings of the 38th International Symposium on Computer Architecture (ISCA ’11). (California, USA)
[3] Mehta N 2015 Xilinx Ultra Scale Architecture for High-Performance, Smarter Systems.
[4] Konotopysen$, KulkovGB, PatrikeevAV, Tarasov I E, Titov A G 2018 Designin of applied VLIS by means of software-based and hardware-based architecture optimizations (in Russ.) In Proc.of Russian conference in Suprecomputing technologies (ST-2018) (Rostov-on-Don: South Federal University) p. 38
[5] Hennessy J L., Patterson D A 2017 Computer Architecture. 6th Edition, A Quantitative Approach. (The Morgan Kaufmann Series in Computer Architecture and Design)
[6] Markovic D, Wang C C, Alarcon L P, Liu T T, Rabaey J M 2010 Ultralow-power design in near-threshold region Proceedings of the IEEE 98 237
[7] Rashid M, Muhtaroglu A 2017 Power delay product optimized hybrid full adder circuits In Poc. of International Artificial Intelligence and Data Processing Symposium (IDAP) (Malatya)
[8] Katreepalli R, Meruguboina D, Haniotakis T 2017 A power-delay efficient carry select adder In
Proc. of 2nd International Conference for Convergence in Technology (I2CT) (Mumbai) pp. 1234-1238.

[9] Praveen P, Zahid A 2017 Variation of power and delay in digital CMOS circuit design in DSM Technology International Journal of Engineering Trends and Technology 45 449

[10] Neve A, Schettler H, Ludwig T, Flandre D 2004 Power-delay product minimization in high-performance 64-bit carry-select adders IEEE Transactions on VLSI systems 12 235