ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 004.436.2:004.75:004.8:004.274:621.382 **doi: 10.32620/reks.2024.3.10**

Artem PEREPELITSYN

National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine

METHODOLOGY OF DEPLOYMENT OF DEPENDABLE FPGA-BASED ARTIFICIAL INTELLIGENCE AS A SERVICE

The subject *of study in this article is the models, methods, and principles of organization of entire lifecycle of Artificial Intelligence (AI) as a Service implemented with the use of Field Programmable Gate Array (FPGA).* ***The purpose*** *of this work is the improvement of the methodology of the deployment of dependable FPGA-based Artificial Intelligence as a Service by creating a complex of mutually agreed concept, principles, models, and methods considering the specifics of the use of heterogeneous computations of Artificial Intelligence and the possibility of realizing the unified protection of FPGA implementations.* ***Tasks****: to clarify the taxonomy of the dependability term within the proposed methodology; to propose the concept of deployment of dependable FPGA-based heterogeneous computations of Artificial Intelligence as a Service; to formulate the principle of tracing changes in FPGA projects and integrated environments during the entire lifecycle; to formulate the principle of unification of protection of FPGA implementations of heterogeneous computing of Artificial Intelligence as a Service; to formulate the principle of the product-service assessment of the availability of FPGA as a Service; and to discuss the promising directions of heterogeneous computations of AI. According to the tasks, the following* ***results*** *were obtained. The existing concepts of dependable systems deployment are discussed. The concept of the deployment of computations of Artificial Intelligence as a Service, which is obtained based on the improvement of paradigms of the creation and deployment of dependable systems and services, is proposed. The principle of tracing of changes, which assumes the updating of requirements during the lifecycle of FPGA projects, is proposed. The principle of the unification of protection, which combines and joins the consideration of various unique features of the FPGA instance to protect the implementation and the set of cyberthreats for the service as a whole, is proposed. The principle of the product-service assessment, which considers parameters and indicators of availability, is proposed. The perspective of the progress of non-electronic mediums for heterogeneous computations with the use of a photonic implementations of Artificial Intelligence computations to ensure improved performance and reduced energy consumption is discussed.* ***Conclusions.*** *One of the main contributions of this research is that in the proposed methodology, the set of principles, models, and methods of deployment of Artificial Intelligence as a Service under conditions of changing requirements and integrated environments, and the need for mechanism of licensing protection of each instance of the system are developed, which allows to reduce model uncertainty by considering various stages of the lifecycle of dependable FPGA implementation using heterogeneous computations.*

Keywords: *dependability of AI; FPGA; tracing of changes; FPGA as a Service; unification of protection; DRM; Artificial Intelligence as a Service; AIaaS; hardware implementation of AI; heterogeneous computing for AI; product-service assessment; QoS; photonic AI.*

Introduction

The progress of FPGA technologies and increasing the level of integration of manufactured devices allows performance rising of hardware systems [1]. It allows to create the services for solving of specialized tasks with high-intensity of computations provided by users [2].

The data processing tasks to support the implementation of elements of Artificial Intelligence are an example of specialized high-intensity computations [3].

The use of heterogeneous computations during the implementation of solutions with elements of AI allows a significant increase of the performance [4] and reducing of energy consumption by means of implementation of the specialized computing tasks [5].

Realization of specialized tasks provided by users with intensive computations required for the implementation of elements of AI [6] in combination with the use of the performance and capabilities of modern FPGA accelerator cards [7] provides the possibility of organizing solutions of AI as a Service [4]. Deep Learning Processor Unit (DPU) simplifies a description of such projects [8]. Improvement of performance of FPGA projects under conditions of continuous changing of requirements and versions of integrated environments is a challenge during the lifecycle of such solutions [9].

Ensuring copyright protection during the deployment of such solutions can be implemented with the use of a licensing mechanism by implementing digital rights management (DRM) [10, 11]. Protecting such solutions from cyberattacks is also the priority direction [12].

Ensuring the reliability of FPGA as a Service and evaluating its parameters [13], including performance and quality of service (QoS) [14], is an important part of the availability assessment of such solutions [15].

Despite hardware implementations of AI in FPGA are in demand [16, 17], there are no ready-made solutions for assessment a set of parameters of such systems.

There is a possibility of using traditional methodological apparatus [18] for AI systems. However, a set of system characteristics does not always allow describing the operation of AI as a Service. But, at the same time, randomly combined components without organization cannot provide such system characteristics.

Thus, it is necessary to find the structure and interactions of the components. In this regard, it is important to create a methodological apparatus and understand the specifics of its application when building such systems.

The main purpose of this research is to improve the methodology of deployment of dependable FPGA-based Artificial Intelligence as a Service by creating a complex of mutually agreed concept, principles, models, and methods considering the specifics of the use of heterogeneous computations of Artificial Intelligence and the possibility of realizing the unified protection of FPGA implementations. To achieve this goal, it is necessary to perform the following **tasks**:

1) to clarify the taxonomy of dependability term within the proposed methodology;

2) to propose the concept of deployment of dependable FPGA-based heterogeneous computations of Artificial Intelligence as a Service;

3) to describe the principle of tracing of changes in FPGA projects and integrated environments during the entire lifecycle;

4) to describe the principle of unification of the protection of FPGA implementations of heterogeneous computing of Artificial Intelligence as a Service;

5) to describe the principle of the product-service assessment of the availability of FPGA as a Service;

6) to discuss promising directions of heterogeneous AI computations.

The structure of this article includes five main sections. The first section defines the understanding of the term dependability within the work. Second section provides a description of the proposed methodology and an explanation of the meaning of the connections between the models, methods, and principles. In the next three sections, the formulation of the three proposed principles with a detailed explanation of their purpose and unique features is provided. Discussion of AI heterogeneous computations is provided in the last section.

1. Taxonomy of dependability within proposed methodology

Taxonomy of dependability implies uncovering the structure of this property. The original definition of dependability assumes a set of attributes, such as reliability, availability, safety, integrity, maintainability, and resulting from them [19]. The evolution of understanding of this term as part of the study of methods for their provision in scientific schools of reliability made it possible to apply it to WEB services [20] and append understanding of meaning of this term [21]. Practical experience demonstrates the need to adapt the meaning of dependability to specific cases [22].

The value set of such a term does not necessarily include the entire range of values. It can be limited to a set of properties that are important and necessary in the case under consideration to convey meaning and significance. Sometimes in publications, it is possible to find reliable and secure or reliable and safe. Reliability corresponds to fault-tolerance. In this case, a distinctive feature is the understanding of the taxonomy of the term dependability for the direction of FPGA-based Artificial Intelligence as a Service.

The reliability aspect considers reliability as part of dependability. The proposed methodology has a fault tolerance aspect, thus, the meaning of reliability is presented in the meaning of term of dependability. The evaluation model of the quality of service includes the reliability aspect, therefore, reliability is included in the set of meaning of term.

The security aspect is also an integral part of dependability. When deploying FPGA-based services, an important issue is the ability to implement a licensing mechanism that relies on cybersecurity at all levels of the service. Therefore, the set of dependability term meaning within the proposed methodology includes the security term because there is the protection of service.

The safety aspect is not considered as part of the meaning of the term dependability within the proposed methodology. It can be an integral part when the cloud infrastructure includes an integral part of safety-critical systems.

Examples of such systems include composite smart homes and smart infrastructure systems, if part of such a system is located in a private cloud. In addition, a list of examples of such systems should include monitoring systems or technical inspection systems for critical infrastructure facilities. In these examples, the safety dimension of the term meaning appears indirectly as an integral part of the concept of term dependability. For an FPGA-based service, this is true if the service performs control functions. However, for solutions that perform analysis tasks, the safety component of the term meaning must be considered separately.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

2. Concept of dependable FPGA-based Artificial Intelligence as a Service

The John von Neumann paradigm of creating reliable systems from unreliable components [23, 24] using redundancy is a primary source for the concept of dependability itself [22].

Further technological improvements require consideration of failures at different levels of the system, and design defects in software components. This brings the evolution of John von Neumann paradigm by other researchers, including those from the scientific school of critical and dependable computing of Doctor of Science on Engineering, Vyacheslav Kharchenko.

The generalization of the dependability taxonomy, as well as the fundamental improvement of the von Neumann concept in the concept of creating dependable systems from non-dependable components, is carried out in the works of Vyacheslav Kharchenko [21, 25].

The component of functional safety in the understanding of dependability is the main component in the proposed methodology for constructing safe systems and infrastructures from insufficiently safe components in the works of Vladimir Sklyar [21].

The concept of creating reliable component-integrated service-oriented systems from non-reliable WEB components with uncertain characteristics is proposed in the works of Anatoliy Gorbenko [20, 26].

The improvement of the von Neumann paradigm not only for the class of systems without maintenance, or for systems with restricted maintenance regulations, but also for systems with maintenance is proposed within the concept of the creation of reliable and secure systems from insufficiently dependable components and multi-purpose maintenance according to combined strategies in the context of changing requirements and the environment of their operation in the works of Yuriy Ponochovnyi [22].

The use of the von Neumann paradigm during the prototyping of Artificial Intelligence services is possible when considering implementation with taking into account into account the features of the components. This principle can be involved because it is possible to perform decomposing into two types of actions. On the one hand, it is necessary to provide a dependable product to allow users to obtain reliable service. On the other hand, this specific service itself is provided as a product. In order to implement such service as dependable, the corresponding characteristics of the product and the service part must be ensured.

In this case, the idea of involving the consideration of a two-component system can be used. At the same time, it is necessary to ensure the functioning of the system, which can be achieved by adopting redundancy measures.

At the top level of such services, redundancy is difficult to implement for such systems. Therefore, in such cases, other solutions can be used to ensure and improve reliability.

Within the proposed concept of deployment of dependable FPGA-based heterogeneous computations of Artificial Intelligence as a Service, it is not easy to use the von Neumann paradigm directly due to the specifics of FPGA-based Artificial Intelligence and the technical complexity of redundancy in the operating service itself.

Therefore, despite the great popularity of the von Neumann paradigm and its application in a number of cases, including the works [22], [23], and [24], in the current implementation, the use of some types of redundancy at the level of the entire service is not practically applicable due to the multiplying costs and expenses. At the same time, the service itself may be not only exclusively one product. Practical implementations may, among other things, be in a different form.

The proposed concept is based on the following statements.

1. FPGA-based Artificial Intelligence as a Service is created, deployed, and supported in the context of changing of target FPGA product ranges, changing project requirements, and versions of integrated development environments during a significant part of the lifecycle.

2. Created implementations of Artificial Intelligence as a Service can be deployed for use both in data centers and on the equipment of the end user with support of heterogeneous computations and provided, including on a subscription basis, which requires ensuring the protection of solutions from unauthorized use.

3. For the end user, a working service is presented as a ready-to-use product that operates under conditions of failures, faults, and changing demand loads, which requires a comprehensive assessment of the availability of such solutions.

Thus, the proposed concept assumes the deployment of FPGA-based Artificial Intelligence as a Service in the context of changing requirements and versions of integrated environments, the need to ensure protection against the unauthorized running of individual instances and availability assessment.

In this case, this is the deployment of a reliable system. In this context the deployment term means more than only the process of installation and includes a wider part of the lifecycle.

The deployment term meaning within the proposed concept covers the lifecycle of such services and implies deployment in a broad sense, including creation.

The proposed concept is based on three main principles. The conceptual relationships between the elements of the proposed methodology are shown in Fig. 1.

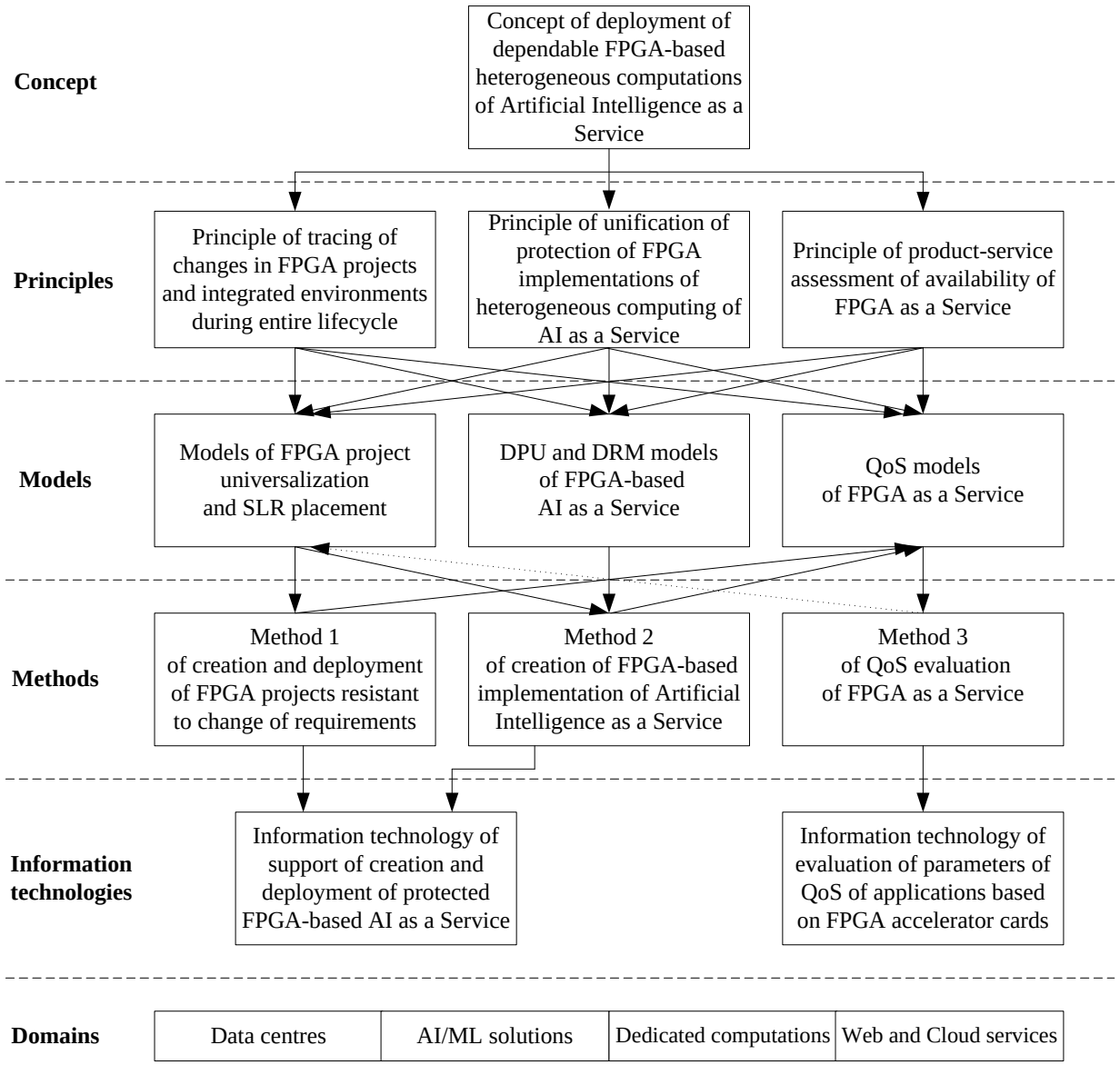

Fig. 1. Proposed elements of methodology of deployment of dependable FPGA-based Artificial Intelligence as a Service

The models are presented in three groups of models, corresponding to the description of FPGA projects, the construction and protection of Artificial Intelligence as a Service, and comprehensive assessments of the availability indicators.

Universalization models of FPGA projects describe the implementation in a form that assumes the possibility of modification of requirements during a significant part of the lifecycle [9, 27].

Organization models of DPU and DRM describe the FPGA Artificial Intelligence project and its protection from the unauthorized use within the service. These models are conceptually presented in publications [3, 4] and in other works [10, 11].

Service assessment models, including QoS, are designed to assess the reliability of project components and the availability of FPGAs as a Service[13, 14].

Methods within the proposed concept of deployment of dependable FPGA-based heterogeneous computations of Artificial Intelligence as a Service presented by a group of complementary methods.

Method of creation and deployment of FPGA projects resistant to change of requirements [27] provides the ability to reduce labor costs during creation and maintaining of projects under changing conditions.

Method of creation of FPGA-based implementation of Artificial Intelligence as a Service [4] allows to create a service with optimization of elements of Artificial Intelligence with hardware implementation, including the FPGA technology.

Method of QoS evaluation of FPGA as a Service provides a comprehensive assessment and the set of steps to reduce the delays of the implementation of FPGA as a Service [14].

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

The relations between models and methods within the proposed concept define their interaction. The dotted arrow from method 3 to the first group of change tracking models in the projects represents the feedback connection. The presence of such feedback connection allows to loop the entire process. In this case, the result of the third QoS assessment method is the input for the project universalization model to take into account the results. In most cases, the connections exist from models to methods, however, in this case, it is a feedback loop to consider the results within the methodology.

Information technology of support of creation and deployment of protected FPGA-based AI as a Service is formed as a result of the union considering the change of projects during the lifecycle and their creation, as well as the creation of AI as a Service. This is one information technology because it considers processes related to each other. Within the first technology, the understanding of the protected term corresponds to the description of the deployment of protected services with licensing with the use of DRM mechanisms.

Information technology of evaluation of parameters of QoS of applications based on FPGA accelerator cards is presented independently and provides the implementation of the assessment that considers a set of parameters related to the elements and reliability indicators included in the models, as well as other parameters. This second information technology takes into account indicators related to understanding of the term dependability within the proposed concept.

3. Principle of tracing of changes in FPGA projects and integrated environments during entire lifecycle

The principle of tracing changes is a continuation of the principle of considering changes in the information and control systems and the environments during the lifecycle [22].

This principle provides changes tracing within the entire FPGA service project, from the versions of FPGA chips, the versions of the integrated environments, and up to the versions of the frameworks that are involved in the interaction process between the FPGA accelerator card and the host computer [28].

Within this principle, the tracing term assumes the movement along the lifecycle of FPGA projects. When moving to development with the use of new integrated environments, it is necessary to change the paradigm of understanding the role of FPGA in the process the creation of the system.

The dynamics of changes of the integrated environments should be considered during the planning of FPGA projects to reduce the risks of possible complete redesign or a completely new prototyping process.

During the programming of solutions for creation of FPGA as a Service, as well as developing and testing the operation of components of Artificial Intelligence systems, it is necessary to consider the transfer of settings from one version of the integrated environment to a newer one when changing or updating the versions.

An example of such a change of versions is the transition from SDAcccel [29] to Vitis [30] from Xilinx.

To unify projects, it is preferable to use a command line interface for processes of work with the project, including the creation of kernels, creation of settings, and compiling and assembling [31].

The unification and possibility of modification of the components of the FPGA project of service, as well as the possibility of transferring and porting such solutions to updated or modified integrated environments, allows to reduce the labor costs for modifying the project in accordance with updated requirements when switching to a new version of the FPGA accelerator card, a version of the framework, or a new version of the integrated environment, as well as during continuous modification of the requirements for such systems.

In this case, it is necessary to follow the recommendations of manufacturer regarding the amount of resources of a specific FPGA package (FPGA housing) and the number of silicon wafers in the chip [32].

This will allow maintaining performance indicators when changing the versions of software tools, FPGA accelerator cards, and after updates of projects.

4. Principle of unification of protection of FPGA implementations of heterogeneous computing of AI as a Service

Within the principle of unified protection it is implied that protection in the broad sense and direct licensing protection are combined. Such protection implies the possibility of accounting for running tasks and operation of a service instance under a license. First of all, this is the possibility of protecting of copyright elements for a specific solution in FPGA, including AI as a Service.

This approach moves forward and expands these two interesting and not entirely directly related aspects of protection. The meaning of the protection term in this principle primarily means ensuring the possibility of licensing with the use of DRM.

At the same time, cybersecurity as an element of information protection is also included because similar tools and cryptographic primitives are used. Within the implementation of DRM for the instance of a service, it is assumed that the use of the same crypto-primitives and embedded solutions built into the kernels in the FPGA which provide protection [10].

The unification term assumes and implies the unification of these two types of FPGA-based Artificial Intelligence as a Service protection in one joint solution. Such a joint solution not only shows the possibility of a task of protection itself but also directly indicates how it is ensured.

In this case, the special property combines cybersecurity and copyright protection. That is, the property of combining copyright protection and information protection.

The distinctive feature and novelty of this principle are not only the indication of the objects themselves that need to be protected but also the proposed principle of unification of protection.

Heterogeneous computing for building implementations with elements of Artificial Intelligence and for their acceleration is very promising. The heterogeneous term within this principle assumes heterogeneous computations but not the heterogeneous systems.

This implies a different form of implementation of the computations. For such systems, computations can be processor-based, based on graphic accelerators, FPGA-based, based on the specialized integrated circuits, and with the use of the implementations in other environments and mediums, including the photonic implementations. Such non-electronic mediums can be used for the acceleration of the computations [33, 34].

5. Principle of product-service assessment of availability of FPGA as a Service

The principle of product-service assessment assumes consideration of all factors affecting QoS.

Within this principle, the considered factors include cybersecurity and technical factors. Thus, hardware failures and faults at different levels of the service organization are also considered as such factors [13].

This assessment is considered at the model level, including the delay assessment [14].

The consideration of fault tolerance of component nodes in FPGA and application of on-chip redundancy, as well as assessment of indicators of fault tolerance in such case are integral part of such assessment [13, 14].

Considering these two components during the assessment in the pure form corresponds to understanding of the dependability term within the taxonomy of this term in this work.

It is proposed to take into account the set of parameters of FPGA-based Artificial Intelligence as a Service.

In contrast to the well-known examples of consideration of the dependability of systems, a distinctive feature of the principle is the consideration of combined details of both the first and second groups of factors with taking into account the specifics of the object of consideration, including different levels of FPGA-based AI as a Service.

A distinctive feature of the deployment of this service is the ability to implement both a project for FPGA accelerator cards and a host application that runs on the host computer to which this set of FPGA accelerator cards is connected.

Therefore, the service is an integral part of the product. They include a working computer, which can be a part of a server stand in a data center, and a set of hardware FPGA accelerators.

This means that, in this case, both the service itself and the product as a whole require attention.

Within this understanding, the service is part of the product. The unification of these two components is implied within this principle of integration and evaluation because both the product and service parts are taken into account.

The consideration may also include the data center infrastructure or the location where the service components operate, or this component may be granted to the provider.

Infrastructural part can be considered as an additional option. In this case, the infrastructure can be considered as the third independent component within the principle. In this case, there are three components: the product, the infrastructure, and the service itself, as a chain of processes (Fig. 2).

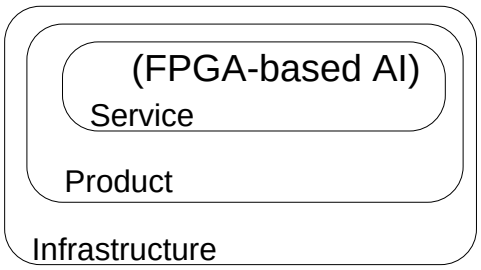

Fig. 2. Structure of consideration of product-service with infrastructure

Only product and service are considered if there is no possibility of influence to the infrastructural part. If, during the creation and deployment of the service, it is possible to use of some existing technical elements provided by the provider, this can be considered as infrastructure. In this case, the electricity, the Internet connection and the rest important conditions for the organization of the service are the responsibility of the provider. It can be assumed that there is no influence on it, and the processes of maintenance are handled by the provider.

In this case, the infrastructure can be excluded from consideration. In this case, only the product and service are considered.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

6. Discussion and implementation

Current implementations of AI systems based on FPGA accelerator cards demonstrate high performance due to the high degree of integration and the use of high-speed dynamic memory, such as High Bandwidth Memory (HBM), with direct interconnects in a single chip package (chip housing). This is an example of heterogeneous computations. The transition to the use of such interconnections instead of printed circuit boards will make the devices even more compact and resistant to external influences [35].

The integration of FPGA and photonic computing environments in a single package (housing) will enable specialized heterogeneous computing based on different environments to solve problems with greater energy efficiency and performance [36].

The use of optical computing to implement the tasks of constructing elements of Artificial Intelligence and services with such computations and functions is extremely promising. In this case, the cost of process of computations can be significantly reduced, and their speed increases up to the speed of signal propagation.

The proposed methodology combines a set of new models and methods, as well as the definition of parameters of dependable FPGA-based Artificial Intelligence as a Service.

Within the proposed principles, it is assumed that changes in the project and the requirements are traced, which represents movement along the lifecycle. The principle of unification of protection represents an expansion and movement taking into account two not entirely directly related aspects of protection, and the principle of product-service assessment of availability involves reviewing the product parts and assessment of availability.

The taxonomy of dependability term within the context of the proposed concept is considered, with the highlighting of the combination of reliable and security terms.

The possibility of creation and deployment of FPGA-based Artificial Intelligence as a Service allows to apply the methodological apparatus of hardware and services for assessment of the parameters.

There are possibilities to model and describe such solutions using such models.

Since this is a service, the model of operation can be presented in the form of a queue of queries. This means that the models of service can be applied. Thus, the Erlang model can be used as a basis.

In this case, FPGA-based AI as a Service differs from non-FPGA-based AI by speed characteristics and a set of delays in constituent units and components.

In this case, it becomes possible to select the parameters. In addition, it is possible to consider failures of the hardware components of the service, as well as possible cyberattacks, and possible problems with the software involved in the hardware implementation of the AI service. These directions are common for all implementations of FPGA as a Service. Model detailing of FPGA-based Artificial Intelligence as a Service requires considering the specifics of such services and implementations.

In addition, the detailing of models involves considering the features of the technological stack used during the creation of such solutions.

This is important for solving of the problem of realization functional reliability model.

In terms of the structural and methodological synthesis of the model, the technological features of the technological stack of modern FPGA solutions, the prototyping tools, and hardware of FPGA accelerator cards are taken into account. The proposed methodology considers such solutions.

Conclusions

The main result of this research is that in the proposed methodology, the set of principles, models, and methods of deployment of Artificial Intelligence as a Service under conditions of changing requirements and integrated environments, and the need of mechanism of licensing protection of each instance of the system, are developed, which allows to reduce model uncertainty with considering of various stages of the lifecycle of dependable FPGA implementation with the use of heterogeneous computations.

The methodology is presented at the system level as a combination of the concept of deployment of dependable FPGA-based heterogeneous computations of Artificial Intelligence as a Service and three proposed principles: the principle of tracing of changes in FPGA projects and integrated environments during the entire lifecycle, the principle of protection unification of FPGA implementations of heterogeneous computing of AI as a Service, and the principle of product-service assessment of the availability of FPGA as a Service.

Thus, the proposed concept allows describing at the system level the process of deployment of dependable projects of FPGA-based Artificial Intelligence as a Service.

Further research can be focused on the improvement of such model with the detailed considering of the features and specifics of the FPGA nature itself and the impact that this technology has on the services that are built on the basis of this technology.

It is also interesting to represent the entire process of the work in the form of a simple model using the Erlang model with the addition of cybersecurity and reliability aspects.

The next step might be the adding horizontal links at the level of principles and the level of models, with specifying of action for each connection. This would allow the set of connections to be viewed not only as a diagram but also as an element of a taxonomic model, allowing the concept to be described in greater detail.

Conflict of interest

The author declares that he has no conflict of interest concerning this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Data availability

The manuscript contains no associated data.

Use of Artificial Intelligence

The author confirms that he did not use any generative technologies of Artificial Intelligence in the work.

The author has read and agreed to the publication of the finale version of the manuscript.

References

1. Chi, T.-K. & et al. An Edge Computing System with AMD Xilinx FPGA AI Customer Platform for Advanced Driver Assistance System. *Sensors*, 2024, vol. 24, iss. 10, no. 3098. DOI: 10.3390/s24103098.

2. Perepelitsyn, A., Zarizenko, I. & Kulanov, V. FPGA as a Service Solutions Development Strategy. *Proceedings 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies, DESSERT 2020*, 2020, pp. 376-380, DOI: 10.1109/DESSERT50317.2020.9125017.

3. Perepelitsyn, A., Fesenko, H., Kasapien, Y. & Kharchenko, V. Technological Stack for Implementation of AI as a Service based on Hardware Accelerators. *Proceedings 2022 IEEE 12th International Conference on Dependable Systems, Services and Technologies, DESSERT 2022*, 2022, pp. 1-5. DOI: 10.1109/DESSERT58054.2022.10018615.

4. Perepelitsyn, A. Method of creation of FPGA based implementation of Artificial Intelligence as a Service. *Radioelectronic and Computer Systems*, 2023, no. 3, pp. 27–36. DOI: 10.32620/reks.2023.3.03.

5. Chun, C. -K. & Lai, K. -C. A Load Balance Scheduling Approach for Generative AI on Cloud-Native Environments with Heterogeneous Resources. *Proceedings 2024 IEEE 10th International Conference on Applied System Innovation, ICASI 2024*, 2024, pp. 223-225, DOI: 10.1109/ICASI60819.2024.10547947.

6. Kalapothas, S., Flamis, G. & Kitsos, P. Efficient Edge-AI Application Deployment for FPGAs. *Information*, 2022, vol. 13, iss. 6, no. 279. DOI: 10.3390/info13060279.

7. *Alveo Product Selection Guide, Data Center Accelerator Cards, Xilinx.* Available at: https://www.xilinx.com/content/dam/xilinx/support/documents/selection-guides/alveo-product-selection-guide.pdf. (accessed July 24, 2024).

8. *Zynq DPU Product Guide, Xilinx, PG338 (v3.3).* Available at: https://docs.xilinx.com/r/3.3-English/pg338-dpu. (accessed February 28, 2023).

9. Perepelitsyn, A. & Kulanov, V. Technologies of FPGA-based projects Development Under Ever-changing Conditions, Platform Constraints, and Time-to-Market Pressure. *Proceedings 2022 IEEE 12th International Conference on Dependable Systems, Services and Technologies, DESSERT 2022*, 2022, pp. 1-5, DOI: 10.1109/DESSERT58054.2022.10018828.

10. Perepelitsyn, A., & Kulanov, V. Analysis of Ways of Digital Rights Management for FPGA-as-a-Service for AI-Based Solutions. *Proceedings 2023 IEEE 13th International Conference on Dependable Systems, Services and Technologies, DESSERT 2023*, 2023. pp. 1-5. DOI: 10.1109/DESSERT61349.2023.10416526.

11. IEEE Recommended Practice for Encryption and Management of Electronic Design Intellectual Property (IP). *in IEEE Std 1735-2014 (Incorporates IEEE Std 1735-2014/Cor 1-2015)*, pp.1-90, 2015, DOI: 10.1109/IEEESTD.2015.7274481.

12. Tetskyi, A. Testuvannia na pronyknennia komponentiv FPGA yak servisu dlia zabezpechennia kiberbezpeky [Penetration testing of FPGA as a Service components for ensuring cybersecurity]. *Aviacijno-kosmicna tehnika i tehnologia – Aerospace technic and technology*, 2023, no. 6, pp. 95–101. DOI: 10.32620/aktt.2023.6.11. (In Ukrainian).

13. Kolesnyk, I., Kulanov, V., Perepelitsyn, A. Markov model of FPGA resources as a service considering hardware failures. *Proc. PhD Symposium at ICTERI 2018*, Kyiv, Ukraine, May 14-17, 2018, CEUR-WS, vol. 2122, pp. 56-62.

14. Perepelitsyn, A., Kulanov, V. & Zarizenko, I. Method of QoS evaluation of FPGA as a service. *Radioelectronic and Computer Systems*, 2022, no. 4, pp. 153–160. DOI: 10.32620/reks.2022.4.12.

15. Shaker, M.N., Hussien, A., Alkady, G.I., Amer, H.H. & Adly, I. FPGA-Based Reliable Fault Secure Design for Protection against Single and Multiple Soft Errors. *Electronics*, 2020, vol. 9, iss. 12, no. 2064. DOI: 10.3390/electronics9122064.

16. Gowda, K.M.V., Madhavan, S., Rinaldi, S., Divakarachari, P.B. & Atmakur, A. FPGA-Based Reconfigurable Convolutional Neural Network Accelerator Using Sparse and Convolutional Optimization. *Electronics* 2022, vol. 11, iss. 10, no. 1653. DOI: 10.3390/electronics11101653.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

17. Seng, K.P., Lee, P.J. & Ang, L.M. Embedded Intelligence on FPGA: Survey, Applications and Challenges. Electronics 2021, vol. 10, iss. 8, no. 895. DOI: 10.3390/electronics10080895.

18. Mehdi, I., Boudi, E.M. & Mehdi, M.A. Reliability, Availability, and Maintainability Assessment of a Mechatronic System Based on Timed Colored Petri Nets. *Appl. Sci*. 2024, vol. 14, iss. 11, no. 4852. DOI: 10.3390/app14114852.

19. Avizienis, A., Laprie, J., Randell, B. & Landwehr, C. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing*, 2004, vol. 1(1), pp. 11-33. DOI: 10.1109/TDSC.2004.2.

20. Gorbenko, A., Kharchenko, V. & Romanovsky, A. On composing Dependable Web Services using undependable web components. *International Journal of Simulation and Process Modelling*, 2007, vol. 3(1/2), pp. 45-54. DOI: 10.1504/IJSPM.2007.014714.

21. Kharchenko, V., Sklyar, V. & Siora, A. Dependability of Safety-Critical Computer Systems through Component-Based Evolution. *2009 Fourth International Conference on Dependability of Computer Systems*, 2009, pp. 42-49. DOI: 10.1109/DepCoS-RELCOMEX.2009.22.

22. Ponochovnyi, Yu. & Kharchenko, V. Metodolohiya zabezpechennya harantozdatnosti informatsiynokeruyuchykh system z vykorystannyam bahatotsil'ovykh stratehiy obsluhovuvannya [Dependability assurance methodology of information and control systems using multipurpose service strategies]. *Radioelektronni i komp'uterni sistemi – Radioelectronic and Computer Systems*, 2020, no. 3, pp. 43–58. DOI: 10.32620/reks.2020.3.05. (In Ukrainian).

23. Beiu, V., Drăgoi, V. -F. & Beiu, R. -M. Why Reliability for Computing Needs Rethinking. *Proceedings 2020 IEEE International Conference on Rebooting Computing ICRC 2020*, 2020, pp. 16-25, DOI: 10.1109/ICRC2020.2020.00006.

24. Palem, K., Lingamneni, A., Enz, C. & Piguet, C. Why design reliable chips when faulty ones are even better. *2013 Proceedings of the ESSCIRC 2013*, 2013, pp. 255-258, DOI: 10.1109/ESSCIRC.2013.6649121.

25. Kharchenko, V. S. Harantozdatni systemy ta bahatoversiyni obchyslennya: aspekty evolyutsiyi [Dependable systems and multiversion computing: aspects of evolution]. *Radioelektronni i komp'uterni sistemi – Radioelectronic and computer systems*. 2009, vol. 7, pp. 46-59.

26. Gorbenko, A., Kharchenko, V., Popov, P. & Romanovsky, A. Dependable Composite Web Services with Components Upgraded Online. *In: de Lemos, R., Gacek, C., Romanovsky, A. (eds). Architecting Dependable Systems III. Lecture Notes in Computer Science*, 2005, vol. 3549, pp. 92–121. DOI: 10.1007/11556169_5.

27. Perepelitsyn, A., & Kulanov, V. Metod stvorennya i vprovadzhennya FPGA proyektiv stiykykh do zmin vymoh i seredovyshch rozroblennya dlya khmarnykh infrastruktur [Method of creation and deployment of FPGA projects resistant to change of requirements and development environments for cloud infrastructures]. *Aviacijno-kosmicna tehnika i tehnologia – Aerospace technic and technology*, 2023, no. 5, pp. 87–97. DOI: 10.32620/aktt.2023.5.07. (In Ukrainian).

28. *Vitis High-Level Synthesis User Guide, AMD, UG1399 (v2024.1).* Available at: https://docs.amd.com/r/en-US/ug1399-vitis-hls/Combining-the-Three-Paradigms. (accessed July 24, 2024).

29. *SDAccel Environment User Guide, Xilinx, UG1023 (v2019.1).* Available at: https://www.xilinx.com/support/documents/sw_manuals/xilinx2019_1/ug1023-sdaccel-user-guide.pdf. (accessed July 24, 2024).

30. *Vitis Unified Software Platform Documentation: Application Acceleration Development, AMD, UG1393 (v2024.1).* Available at: https://docs.amd.com/r/en-US/ug1393-vitis-application-acceleration/Getting-Started-with-Vitis. (accessed July 24, 2024).

31. *Vitis Unified IDE and Common Command-Line Reference Manual, AMD, UG1553 (v2023.1).* Available at: https://docs.amd.com/r/en-US/ug1553-vitis-ide. (accessed July 24, 2024).

32. *UltraFast Design Methodology Guide for Xilinx FPGAs and SoCs, Xilinx, UG949 (v2021.2).* Available at: https://docs.xilinx.com/r/2021.2-English/ug949-vivado-design-methodology/SLR-Utilization-Considerations. (accessed February 28, 2023).

33. Ban, Y. Silicon Photonics for Scaling the Cloud and Enabling AI. *Proceedings 2022 IEEE International Symposium on VLSI Design, Automation and Test, VLSI-DAT 2022*, 2022, pp. 1-1, DOI: 10.1109/VLSI-DAT54769.2022.9768090.

34. Ning, S. & et al. Photonic-Electronic Integrated Circuits for High-Performance Computing and AI Accelerators. *in Journal of Lightwave Technology*. 2024. pp. 1-26. DOI: 10.1109/JLT.2024.3427716.

35. Shakoorzadeh, N. & et al. Reliability Studies of Silicon Interconnect Fabric. *Proceedings 2019 IEEE 69th Electronic Components and Technology Conference, ECTC 2019*, 2019, pp. 800-805, DOI: 10.1109/ECTC.2019.00126.

36. Tossoun, B. & et al. Large-Scale Integrated Photonics for Energy-Efficient AI Hardware. *Proceedings 2024 IEEE Photonics Society Summer Topicals Meeting Series SUM 2024*, 2024, pp. 1-2, DOI: 10.1109/SUM60964.2024.10614557.

Received 15.06.2024, Accepted 20.08.2024

МЕТОДОЛОГІЯ ПОБУДОВИ ГАРАНТОЗДАТНОЇ FPGA РЕАЛІЗАЦІЇ ШТУЧНОГО ІНТЕЛЕКТУ ЯК СЕРВІСУ

А. Є. Перепелицин

Предметом вивчення в даній статті є моделі, методи та принципи побудови з урахуванням всього життєвого циклу штучного інтелекту (ШІ) в якості сервіса та реалізацією з використанням технології FPGA. **Метою** даного дослідження та статті є розвиток методології побудови гарантоздатної FPGA реалізації штучного інтелекту та надання таких рішень в якості сервіса шляхом створення комплексу взаємоузгоджених концепції, принципів, моделей і методів, враховуючи особливості використання гетерогенних обчислень для реалізації проєктів штучного інтелекту як сервісу і можливості реалізації уніфікації захисту FPGA проєктів. **Завдання:** роз'яснити систематику значення поняття гарантоздатність в рамках запропонованої методології; запропонувати концепцію побудови гарантоздатних гетерогенних обчислень штучного інтелекту як сервісу на основі FPGA; сформулювати принцип відстеження змін у проєктах FPGA та інтегрованих середовищах протягом усього життєвого циклу; сформулювати принцип уніфікації захисту FPGA реалізацій гетерогенних обчислень штучного інтелекту в якості сервіса; сформулювати принцип продуктно-сервісного оцінювання готовності FPGA як сервісу; обговорити перспективні напрямки удосконалення гетерогенних обчислень ШІ. Відповідно до поставлених завдань, були отримані наступні **результати**. Обговорюються існуючі концепції побудови гарантоздатних систем. Запропонована концепція побудови обчислень штучного інтелекту як сервісу, яка отримана шляхом розвитку парадигм створення і розгортання гарантоздатних систем і сервісів. Запропоновано принцип відстеження змін, який передбачає оновлення вимог протягом життєвого циклу проєктів FPGA. Запропоновано принцип уніфікації захисту, який об'єднує та поєднує врахування різних видів властивостей екземпляра FPGA для захисту реалізації та множину кіберзагроз для сервіса в цілому. Запропоновано принцип продуктно-сервісного оцінювання, який розглядає показники готовності. Обговорюються перспективи розвитку неелектронних гетерогенних обчислень з використанням фотонних реалізацій, які забезпечують збільшення швидкодії та зниження енергоспоживання обчислень штучного інтелекту. **Висновки:** один із головних внесків цього дослідження полягає в тому, що у рамках запропонованої методології розроблено набір принципів, моделей і методів побудови штучного інтелекту як сервісу в умовах змін вимог, інтегрованих середовищ та необхідності захисту прав власності окремого екземпляра системи, що дозволяє знизити модельну невизначеність з урахуванням різних етапів життєвого циклу гарантоздатної FPGA реалізації з використанням гетерогенних обчислень.

Ключові слова: гарантоздатність ШІ; FPGA; відстеження змін; FPGA як сервіс; уніфікація захисту; DRM; DPU; штучний інтелект як сервіс; AIaaS; апаратна реалізація ШІ; гетерогенні обчислення для ШІ; продуктно-сервісне оцінювання; QoS; фотонний ШІ.

Перепелицин Артем Євгенович – канд. техн. наук, доц., доц. каф. комп'ютерних систем, мереж і кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Artem Perepelitsyn – PhD, Associate Professor at the Computer Systems, Networks and Cybersecurity Department, National Aerospace University «Kharkiv Aviation Institute», Kharkiv, Ukraine,
e-mail: a.perepelitsyn@csn.khai.edu, ORCID: 0000-0002-5463-7889, Scopus Author ID: 56332607800.