ISSN 0002-306X. Proc. of the RA NAS and NPUA Ser. of tech. sc. 2024. V. LXXVII, N1

UDC 621.382 **MICROELECTRONICS**

DOI: 10.53297/0002306X-2024.v77.1-87

A.A. AVETISYAN, M.T. GRIGORYAN, A.V. MELIKYAN

BENCHMARKING AND IMPLEMENTING DEEP LEARNING ALGORITHMS ON FIELD PROGRAMMABLE GATE ARRAYS AND APLLICATION SPECIFIC INTEGRAL CIRCUIT PLATFORMS

Issues on the enhancement of Artificial Intelligence (AI) performance using Field-Programmable Gate Arrays (FPGA) and Application-Specific Integrated Circuits (ASIC) are studied. It focuses on benchmarking and implementing deep learning algorithms, crucial components of modern AI, on these advanced hardware platforms. The study begins with an explanation of the significance of deep learning in AI and the growing need for efficient computing platforms like FPGA and ASIC. These platforms are known for their high-speed processing capabilities and low power consumption, making them ideal for AI applications.

The research then delves into a detailed analysis of how deep learning algorithms can be optimized and executed on FPGA and ASIC platforms. It highlights the methods used to benchmark the performance of these algorithms on the mentioned hardware, providing a clear comparison with traditional computing systems. The paper also discusses the challenges and solutions in integrating deep learning algorithms into these specialized hardware environments.

Further, the advantages of using FPGA and ASIC for AI tasks, including improved processing speed, reduced energy consumption, and enhanced ability to handle complex AI computations are studied.

Keywords: FPGA, ASIC, artificial intelligence, neural networks.

Introduction. The field of artificial intelligence (AI) has seen an unprecedented acceleration in performance and efficiency, primarily driven by significant advancements in deep learning algorithms and their implementation on specialized hardware platforms. This paper delves into the comparative analysis and practical implications of deploying deep learning models on Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs), two of the leading hardware platforms that offer distinct advantages for AI applications. It is predicted that the AI on chip market revenue will be rising exponentially in the coming decade [1] and will increase more than 13 times until 2032 (Fig. 1).

FPGAs, known for their flexibility and reconfigurability, present a compelling option for AI research and development, allowing for rapid prototyping and adaptation to evolving algorithmic needs. The adaptability of FPGAs to changing requirements

and algorithms is discussed in-depth in the [2] works of Hauck and DeHon (2010), who highlight the architectural benefits and design considerations of FPGAs for computing tasks.

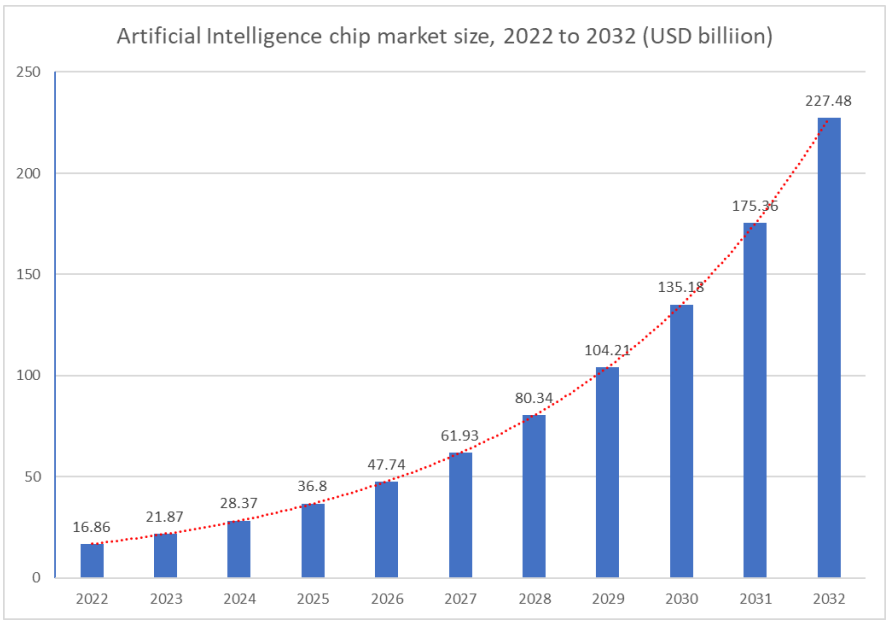

Fig. 1. AI on chip market revenue predictions

On the other hand, ASICs, with their specialized design tailored for specific applications, offer unmatched efficiency and performance for well-defined tasks. [3] explores the design space and optimization strategies for ASICs in AI, emphasizing the potential for achieving high throughput and energy efficiency in deep learning applications.

The core of this paper focuses on benchmarking the performance of deep learning algorithms when implemented on these platforms, considering metrics such as computational throughput, power consumption, and latency. Benchmarking efforts draw on the methodology outlined in [4] comprehensive analysis of deep learning benchmarks across various hardware platforms, providing a framework for evaluating FPGA and ASIC implementations. The data from [4] is summed up in Fig. 2.

By integrating insights from these reference papers, the current study offers a nuanced understanding of the trade-offs involved in choosing between FPGAs and ASICs for AI tasks. It examines how the inherent flexibility of FPGAs might be leveraged for experimental and evolving AI models, while the efficiency of ASICs could be harnessed for large-scale, high-performance applications with stable requirements.

Furthermore, this paper contributes to the ongoing discourse on optimizing hardware architectures for AI by presenting case studies and empirical data on the implementation of cutting-edge deep learning algorithms on both FPGA and ASIC platforms. Through this analysis, it seeks to provide actionable insights for researchers and practitioners in the field of AI, guiding the selection and optimization of hardware platforms for diverse AI applications.

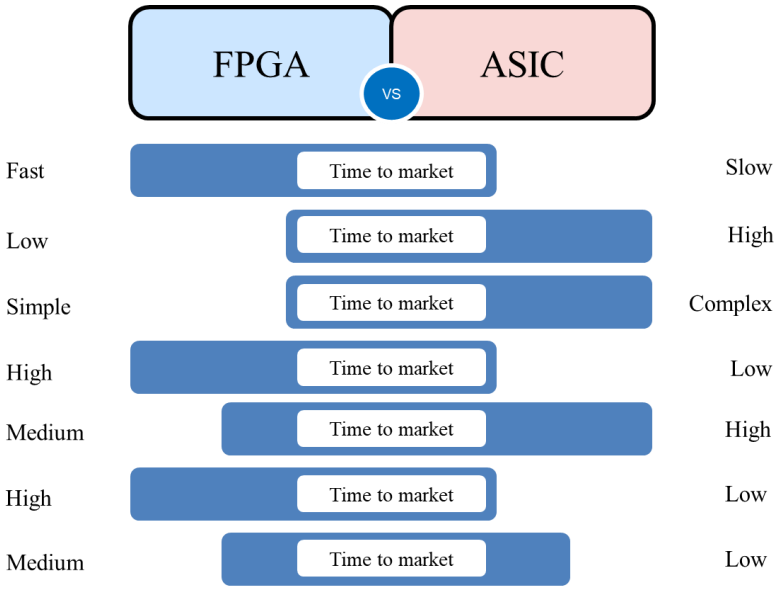

Fig. 2. Pros and cons of FPGA against ASIC

Incorporating an examination of benchmarking methods into the discourse on the implementation of deep learning algorithms on FPGA and ASIC platforms is essential, especially given the inherent challenge of establishing equivalency between disparate FPGA models and ASIC designs. This paper extends its analysis to address these benchmarking intricacies, adopting a multi-faceted approach to navigate the heterogeneity of hardware specifications and performance metrics.

Benchmarking deep learning implementations on FPGAs and ASICs involves a nuanced methodology that accounts for not only raw performance metrics such as computational throughput (in operations per second) and power efficiency (in operations per watt), but also factors like programmability, scalability, and the adaptability of the hardware to evolving deep learning models. The complexity of this task is amplified by the diversity in FPGA architectures and the specificity of ASIC designs, which necessitates a standardized yet flexible benchmarking framework.

In essence, the benchmarking methodology outlined in this paper is designed to provide a fair comparison of FPGA and ASIC platforms for deep learning applications. By addressing the challenge of establishing equivalency between heterogeneous hardware platforms, this approach enables a more informed decision-making process for researchers and practitioners in the field of AI, guiding the selection of the most suitable hardware for specific deep learning tasks and objectives.

Method. To perform proper benchmarking between FPGA device and ASIC design two aspects are compared:

- Power efficiency
- Maximum clock period

Of course, to ensure equivalent conditions for comparison several considerations must be made.

1. The same RTL code is used during synthesis for FPGA and ASIC design.
2. During synthesis and implementation stages, the same constraints are used for FPGA and ASIC design.
3. ASIC design uses a transistor library for the same technology and operating voltage as the FPGA device's internal logic. During the analysis stage both environments are simulated in a typical corner.
4. For both devices, the same generated RTL code is used. The RTL is generated by using several caffe models and executing IBM's AccDNN tool [5].
5. For FPGA, the RTL models are synthesized, implemented, and analyzed using Xilinx Vivado tool [6]. The board used in the research is Vitrix-7.

The flow used to achieve proper benchmarking between ASIC and FPGA is illustrated in Figure 3 which includes all mentioned steps. ASIC design is created and simulated by Synopsys Design Compiler tool [7] (Fig. 4). The technology used for the implementation is Synopsys Advanced Education Design 32/28 nanometer (nm) library, since it is the same one used for Virtix-7 FPGA boards. The transistor library is selected based on the operating voltage (1.05V), process (typical) and the transistor type (rvt). The constraint file between Vivado and Design Compiler tools is shared since both support the *.xdc format.

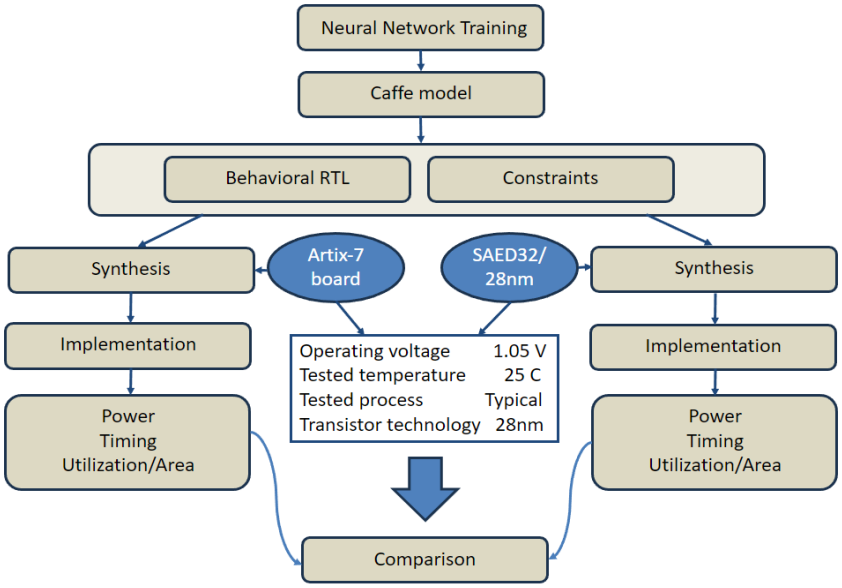

Fig. 3. FPGA vs ASIC benchmarking flow

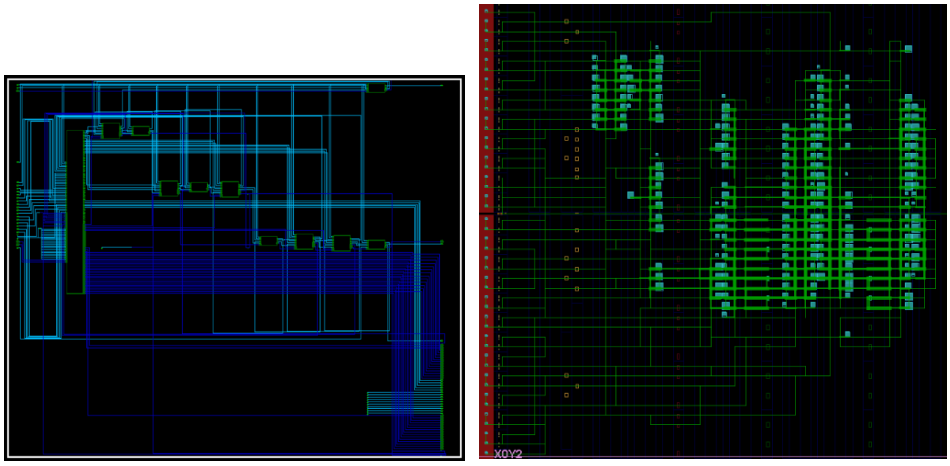

Fig. 4. ASIC design synthesized by Design Compiler (on the left). Design placed on FPGA (on the right)

Experimental results. Several neural networks were passed through the mentioned flow (particularly cifar10 [8], vgg16 [9], yolo [10], Alexnet [11]). Both static and dynamic powers were measured and compared. Total power is equal to the sum of static and dynamic powers as shown in Figure 5.

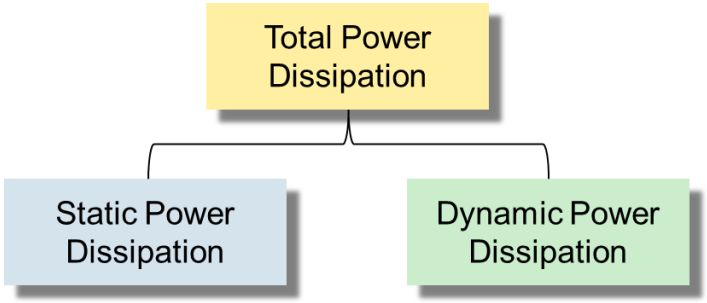

Fig. 5. Power dissipation in ICs

$$S = \int_0^t V_{DD} I_{leak}\, dt, \tag{1}$$

where S is the static power dissipation, V_{DD} – the source voltage, I_{leak} – the total leakage current of the system.

$$D = \int_0^t C V_{DD}^2 f_c dt, \tag{2}$$

where D is the dynamic power dissipation, C – the total switching capacitance of the circuit, V_{DD} – the source voltage, f_c – switching activity or frequency.

Finally, the total power dissipation, which is the sum of D and S will be:

$$E = D + S = \int_0^t (V_{DD} I_{leak} + C V_{DD}^2 f_c) dt. \tag{3}$$

The tools are measuring the system power based on equations (1), (2), (3), and the results are shown in Table I.

As can be seen from the results, all ASIC designs show significantly lower power consumption due especially in the dynamic domain. The main contributors to this reduction are optimized design, lower parasitic due to shorter and more efficient routing and clock paths.

The second aspect of benchmarking is the maximum clock frequency which directly impacts the computation capability of the design. Apart from the device limitations, the only entity which keeps the working frequency down is the slack represented in formula (4):

$$Slack = Arrival\_time - Required\_time, \tag{4}$$

where Arrival_time is the time elapsed for a signal to arrive at a certain point.

Table 1

Power measurement for several ANN implementations (in Watts)

Platform	Power	cifar10	vgg16	yolo	Alexnet
Virtix-7	Dynamic	2,328	7,128	7,338	7,212
	Static	0,244	0,736	0,797	0752
ASIC	Dynamic	0,189	0,4859	0,5132	0,4808
	Static	0,0235	0,0849	0,0927	0,08997

The Required_time is the latest time at which a signal can arrive without making the clock cycle longer than desired.

The slack should always be positive, otherwise the design will not function properly.

Table 2

The maximum working frequency with the acceptable slack for several ANN implementations (in MHz)

Platform	cifar10	vgg16	yolo	Alexnet
Virtix-7	186,7	219,3	128,0	185.4
ASIC	568,4	472.3	232,9	281.9

Conclusion. This study has provided evaluation of deep learning algorithm performance on FPGA and ASIC platforms, highlighting the trade-offs between flexibility and efficiency. The findings demonstrate that while FPGAs offer adaptability and are conducive to research and development, ASICs excel in high-throughput, energy-efficient computations for established deep learning tasks. The benchmarking methodology adopted ensures a fair and informative comparison, considering factors such as power efficiency, clock frequency, and design area. The data shows improved power efficiency of ASIC designs compared to FPGA by average 90.2%. At the same time, ASIC designs have 67.5% higher clock frequency capability. The future work will expand on the implications of these findings for the design of more specialized hardware and the optimization of deep learning algorithms for these platforms. Also, microbenchmarking can be added to the methodology to increase the equivalency between the devices, and gain more insight into the further design optimizations.

REFERENCES

1. Presedence Research Pvt Ltd. Artificial Intelligence (AI) Chip Market Global Industry Analysis, Size, Share, Growth, Trends, Regional Outlook, and Forecast 2023-2032. Available: https://www.precedenceresearch.com/artificial-intelligence-chip-market 2022
2. **Hauck S. and DeHon A.** Reconfigurable Computing: The Theory and Practice of FPGA-Based Computation.- San Francisco, CA, USA: Morgan Kaufmann, 2008.- P. 129-155.
3. **Chen C.H., Knag P., Zhang Z.** Characterization of heavy-ion-induced single-event effects in 65 nm bulk CMOS ASIC test chips// IEEE Transactions on Nuclear Science.- August 2014.-vol. 61, no. 5.
4. **Dai W. and Berleant D.** Benchmarking Contemporary Deep Learning Hardware and Frameworks: A Survey of Qualitative Metrics // 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI).- Los Angeles, CA, USA.- 2019.- P. 148-155.- doi: 10.1109/CogMI48466.2019.00029.
5. DNNBuilder: an Automated Tool for Building High-Performance DNN Hardware Accelerators for FPGAs/**Xiaofan Zhang, Junsong Wang, Chao Zhu, Yonghua Lin, et al** // IEEE Transactions on Very Large Scale Integration (VLSI) Systems.- April 2020.- Vol. 30, no. 4.- P. 123-136.
6. Xilinx, Inc., "Xilinx Vivado Design Suite" 2023. Available: https://www.xilinx.com/products/design-tools/vivado.html. Accessed: Jan. 31, 2024.
7. Synopsys, Inc., "Synopsys Design Compiler", 2021. Available: https://www.synopsys.com/implementation-and-signoff/rtl-synthesis-test/design-compiler.html. Accessed: Jan. 31, 2024.
8. **Krizhevsky A.** Learning multiple layers of features from tiny images: Master's thesis.- Department of Computer Science, University of Toronto, 2009.
9. **Karen S., Andrew Z.** Very Deep Convolutional Networks for Large-Scale Image Recognition.-2014.-arXiv 1409.1556v6.
10. **Joseph Redmon, Ali Farhadi.** YOLOv3: An Incremental Improvement.-University of Washington, 2018.
11. **Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton.** ImageNet Classification with Deep Convolutional Neural Networks.-University of Toronto Canada, 2012.

National Polytechnic University of Armenia. The material is received on 29.02.2024

Ա.Ա. ԱՎԵՏԻՍՅԱՆ, Մ.Տ. ԳՐԻԳՈՐՅԱՆ, Ա.Վ. ՄԵԼԻՔՅԱՆ

ԽՈՐ ՈՒՍՈՒՑՄԱՆ ԱԼԳՈՐԻԹՄԵՐԻ ԹԵՄԱՏԻԿ ԾՐԱԳՐԱՎՈՐՎՈՂ ՓԱԿԱՆՆԵՐԻ ԶԱՆԳՎԱԾԻ ԵՎ ԿԻՐԱՌՈՒԹՅԱՆԸ ԿՈՂՄՆՈՐՈՇՎԱԾ ԻՆՏԵԳՐԱԼ ՍԽԵՄԱՆԵՐԻ՝ ՀԱՐԹԱԿՆԵՐԻ ՎՐԱ ԻՐԱԿԱՆԱՑՈՒՄԸ ԵՎ ՀԱՄԵՄԱՏՈՒՄԸ

Ներկայացվել է արհեստական բանականության (ԱԲ) կատարողականի բարելավման վերաբերյալ ուսումնասիրություն՝ օգտագործելով թեմատիկ ծրագրավորվող փականների զանգվածները (ԹԾՓՋ) և կիրառությանը կողմնորոշված ինտեգրալ սխեմաներ (ԿԿԻՍ): Աշխատանքում ուշադրությունը կենտրոնացվել է այս սարքային հարթակներում խոր ուսուցման ալգորիթմների՝ ժամանակակից ԱԲ-ի կարևոր բաղադրիչների չափորոշման և ներդրման վրա: Ուսումնասիրությունը սկսվում է ԱԲ-ում խոր ուսուցման նշանակության և արդյունավետ հաշվողական հարթակների, ինչպիսիք են ԹԾՓՋ-ն և ԿԿԻՍ-ը, աճող անհրաժեշտության ներկայացմամբ: Այս հարթակները հայտնի են իրենց բարձր արագությամբ մշակվելու հնարավորություններով և էներգիայի ցածր սպառմամբ, ինչը դրանք դարձնում է ԱԲ հավելվածների իրականացման համար գերազանց միջավայր: Հետազոտությամբ մանրամասն վերլուծվում է, թե ինչպես կարող են խոր ուսուցման ալգորիթմները լավարկվել և գործարկվել ԹԾՓՋ և ԿԿԻՍ հարթակներում: Ընդգծվում են նշված սարքավորումների վրա ալգորիթմների կատարողականությունը համեմատելու համար օգտագործվող եղանակները՝ ապահովելով հստակ համեմատություն ավանդական հաշվողական համակարգերի հետ: Քննարկվում են նաև այս սարքային մասնագիտացված միջավայրում խոր ուսուցման ալգորիթմների ներառման մարտահրավերներն ու լուծումները: Հետազոտվում են ԱԲ առաջադրանքների համար ԹԾՓՋ-ի և ԿԿԻՍ-ի օգտագործման առավելությունները, ներառյալ մշակման բարելավված արագությունը, էներգիայի կրճատված սպառումը և բարդ ԱԲ հաշվարկներ կատարելու բարձրացված ունակությունը:

Առանցքային բառեր. թեմատիկ ծրագրավորվող փականների զանգվածներ, կիրառությանը կողմնորոշված ինտեգրալ սխեմաներ, արհեստական բանականություն, նեյրոնային ցանցեր:

А.А. АВЕТИСЯН, М.Т. ГРИГОРЯН, А.В. МЕЛИКЯН

РЕАЛИЗАЦИЯ И СРАВНЕНИЕ АЛГОРИТМОВ ГЛУБОКОГО ОБУЧЕНИЯ НА ПЛАТФОРМАХ ПРОГРАММИРУЕМЫХ ВЕНТИЛЬНЫХ МАТРИЦ И ИНТЕГРАЛЬНЫХ СХЕМ СПЕЦИАЛЬНОГО НАЗНАЧЕНИЯ

Исследуются вопросы повышения производительности искусственного интеллекта (ИИ) с использованием программируемых вентильных матриц (ПВМ) и интегральных схем специального назначения (ИССН). Основное внимание уделяется тестированию и внедрению алгоритмов глубокого обучения, важнейших компонентов современного искусственного интеллекта, на этих передовых аппаратных платформах. Исследование

начинается с объяснения значения глубокого обучения в ИИ и растущей потребности в эффективных вычислительных платформах, таких как ПВМ и ИССН. Эти платформы известны своими возможностями высокоскоростной обработки и низким энергопотреблением, что делает их идеальными для приложений искусственного интеллекта. Затем исследование углубляется в подробный анализ того, как алгоритмы глубокого обучения могут быть оптимизированы и реализованы на платформах ПВМ и ИССН. В нем освещаются методы, используемые для оценки производительности этих алгоритмов на упомянутом оборудовании, обеспечивая четкое сравнение с традиционными вычислительными системами. Обсуждаются проблемы и решения по интеграции алгоритмов глубокого обучения в эти специализированные аппаратные среды. Изучаются преимущества использования ПВМ и ИССН для задач ИИ, включая повышение скорости обработки, снижение энергопотребления и расширение возможностей обработки сложных вычислений ИИ.

Ключевые слова: программируемые вентильные матрицы, интегральные схемы специального назначения, искусственный интеллект, нейронные сети.