Evaluation of heterogeneous AIoT Accelerators within VEDLIoT

R. Griessl*, F. Porrmann*, N. Kucza*, K. Mika*, J. Hagemeyer*, M. Kaiser*, M. Porrmann†, M. Tassemeier†, M. Flottmann†, F. Qararyah‡, M. Waqar‡, P. Trancoso‡, D. Ödman§, K. Gugala¶, G. Latosinski¶
*Bielefeld University, Germany — †Osnabrück University, Germany —
‡Chalmers University of Technology, Sweden — §EMBEDL AB, Sweden — ¶Antmicro, Poland

***Abstract*—Within VEDLIoT, a project targeting the development of energy-efficient Deep Learning for distributed AIoT applications, several accelerator platforms based on technologies like CPUs, embedded GPUs, FPGAs, or specialized ASICs are evaluated. The VEDLIoT approach is based on modular and scalable cognitive IoT hardware platforms. Modular microserver technology enables the integration of different, heterogeneous accelerators into one platform. Benchmarking of the different accelerators takes into account performance, energy efficiency and accuracy. The results in this paper provide a solid overview regarding available accelerator solutions and provide guidance for hardware selection for AIoT applications from far edge to cloud.**

VEDLIoT is an H2020 EU project which started in November 2020. It is currently in an intermediate stage. The focus is on the considerations of the performance and energy efficiency of hardware accelerators. Apart from the hardware and accelerator focus presented in this paper, the project also covers toolchain, security and safety aspects. The resulting technology is tested on a wide range of AIoT applications.

I. Introduction

The VEDLIoT project focuses on energy-efficient Deep Learning (DL) for AIoT (Artificial Intelligence of Things) use cases. Looking into novel architectures, optimized to accelerate the computation of neural networks, VEDLIoT comes up with adaptable and scalable hardware solutions tailored towards the requirements of applications. The overall approach of VEDLIoT covers a fully-featured, heterogeneous hardware platform, integrating the accelerators described and evaluated within this paper. The hardware platform is described in more detail in Section II-A. VEDLIoT also includes aspects regarding supported toolchains, as well as developments regarding security and safety components. A more detailed introduction to the overall project can be found in [1]. VEDLIoT hardware platforms are used to realize use cases in the areas of smart home, industrial IoT as well as automotive. The VEDLIoT Open Call is used to include ten additional use cases from a wide range of AIoT areas, covering the agricultural, healthcare and medical domains.

Over the last years, a large number of diverse DL accelerators in the form of special ASICs or IP cores as well as GPUs- or FPGA-based solutions have been introduced in the market. VEDLIoT has put great effort into the benchmarking and comparable evaluation of selected accelerators regarding performance, energy efficiency and accuracy. Together with the seamless integration of DL into the VEDLIoT IoT hardware platforms, the benchmarking methodology is used for further optimizing applications towards performance and energy efficiency. In this paper, we present a summary of the results obtained. More details are available in the respective project deliverables [2], [3].

II. Hardware Platform and Accelerators

This section deals with the VEDLIoT hardware architecture and presents the different accelerators evaluated. It also acts as an introduction and classification for the different accelerators used in the benchmarking section (Section III).

A. Hardware Platform

VEDLIoT's hardware platforms are a joint infrastructure for the developments within the project. A wide range of AIoT applications can be addressed using a flexible communication infrastructure and exchangeable microservers. Figure 1 shows the RECS platforms covering application domains from embedded/far-edge computing towards cloud computing. All platforms commonly target heterogeneous computing with tightly coupled microservers. The cloud computing platform RECS|Box consists of either two or three rack units and aims for high-density applications using hundreds of microservers with high-bandwidth communication requirements. t.RECS houses up to three microservers in one rack unit and focuses on edge computing scenarios with low-latency demands like VEDLIoT's SmartMirror use-case or 5G base stations in the automotive use-case. u.RECS rounds off the range of the RECS family towards low-power and compact embedded computing.

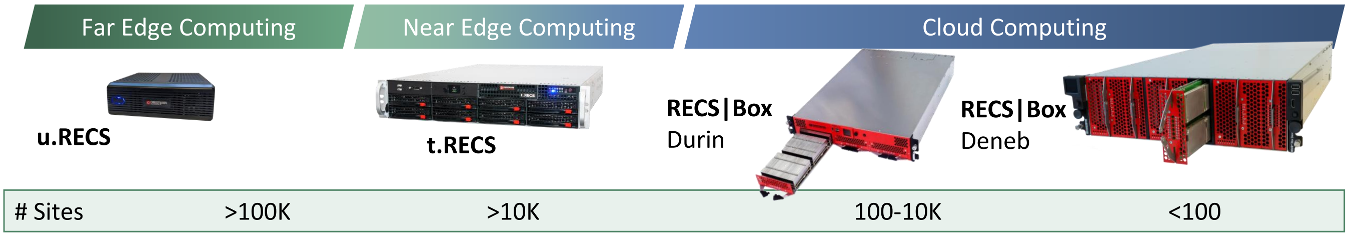
Microservers are based on industry standard Computer-on-Module (COM) form factors, allowing for flexible and heterogeneous processing. RECS|Box and t.RECS support microservers are based on COM Express and COM-HPC Server and Client standards. The u.RECS, on the other hand, supports with SMARC, Jetson NX, Xilinx Kria and Raspberry Pi compute modules multiple compact form factors for far edge computing

B. Accelerator Overview

There are many accelerators available for a wide range of applications, from small embedded systems with power bud-

This publication incorporates results from the VEDLIoT project, which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957197.

978-3-9819263-7-8/DATE23/© 2023 EDAA



# Sites	>100K	>10K	100-10K	<100
Footprint	Custom	Compact (1RU)	Medium (2RU)	Large (3RU)
Power Budget	<30 W	< 500 W	500 W – 2 KW	> 2 KW
# Microserver	max 2	up to 3	up to 48	up to 144

Fig. 1: Overview of modular and scalable RECS platforms

gets in the order of milliwatt to cloud platforms with a power consumption exceeding 400 W. Figure 2 provides an overview of the different accelerators using a double logarithmic plot, grouping them into three groups, depending on their peak performance values (in Giga Operations per Second). It should be noted that values provided by the vendors are used, so no normalization regarding technology, precision or architecture is performed. On average, an energy efficiency of about 1 Tera Operation per W (1 TOPS/W) is achieved. In the following paragraphs, the main characteristics of the three performance groups are discussed.

Ultra Low Power ($<$ 3 W): The ultra-low power group of accelerators are mainly devices integrating energy-efficient, microcontroller-style cores combined with compact accelerators for DL-specific functions. They are focusing on generic IoT applications like the Maxim MAX78000, the Ambient Scientific GPX-10 or the BrainChip Akida, providing only simple analogue or digital interfaces. Other devices such as the Greenwave GAP 8 and GAP 9, the Canaan Kendryte K210 or the Kneron KL530 and KL720 also aim at vision processing, providing an additional camera interface. Typically, those devices are directly designed into the application itself without using a modular or microserver-based approach, simply because all interfaces and peripherals are integrated. Only the Bitmain Sophon BM1880 and Intel Myriad X are providing a generic USB interface and are designed to act as accelerator devices attached to a regular host processor. None of these devices integrate external memory controller interfaces. Based on its wide availability, the Intel Myriad X device is included in the benchmarking activity.

Low Power (3 W to 35 W): While the previous group of accelerators is focusing on applications with a very low power envelope, often in a battery-powered environment with no special requirements regarding cooling, the low-power group of accelerators includes accelerators for a wide range of applications in automation and automotive. All devices include high-speed interfaces for external memories, and peripherals, as well as high-speed communication towards other processing devices or host systems, such as PCIe, proving excellent capabilities for a modular, microserver-based approach as supported by the RECS platform. Apart from the Hailo-8, the FlexLogix InferX X1 and the VSORA Tyr family, which are designed as dedicated accelerators attached to an external host processor, all devices include powerful, general-purpose application processors, capable of running a fully-fledged Linux operating system. In addition to specialized ASICs including the Coherent Logix HX40416, the Blaize El Cano or the Huawei Ascend 310, this group also includes embedded

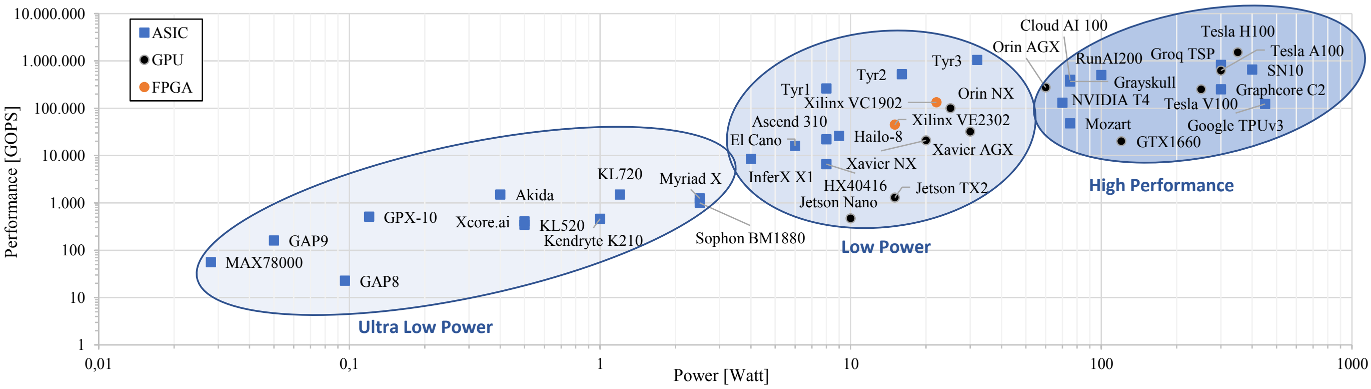


Fig. 2: Peak performance of AI Accelerators and classification into performance groups

GPUs from NVIDIA, in particular the Jetson family, starting from the Nano and TX2, via the Xavier NX and Orin NX devices all the way up to the AGX Xavier. The Xilinx Versal Core AI VC1902 and Versal Edge AI VE2302 are explained in detail in Section II-C.

High Performance ($> 35\,\mathrm{W}$): The high-performance group of accelerators includes devices with up to 450 W of TDP, suitable for both inference and training use-cases, typically deployed in the form of a PCIe extension cards for edge or cloud servers. Besides the classical NVIDIA Tesla GPGPUs including Tesla V100, A100 and H100, also dedicated ASICs like the Groq TSP, the SambaNova SN10, the Graphcore C2 or the Google TPUv3 are part of this cluster. In addition, also powerful inference ASICs like the SimpleMachines Mozart, the Tenstorrent Grayskull, the Qualcomm Cloud AI 100 Chip or the Untether AI RunAI200 are included. As a reference, also a consumer-class NVIDIA Geforce GTX 1660 GPU has been included in the benchmarking. The NVIDIA Jetson AGX Orin is also part of this group due to its high power envelope, although it's part of the embedded NVIDIA Jetson family.

C. Reconfigurable Accelerators

In addition to the integration of the latest off-the-shelf accelerators for deep learning into the RECS environment, VEDLIoT targets the development of customized accelerators for dedicated DL algorithms using FPGAs. As a baseline for later comparison to the own developments, the Xilinx Deep Learning Processor Unit (DPU) is used. Additionally, the DPU is used as a basis for a dynamically reconfigured accelerator, which can be adapted to different application requirements at runtime.

An FPGA base design has been developed to support the RECS platform with its flexible internal and external interconnect and to ease integration of new FPGA-based accelerators. The FPGA architecture contains a block-based design for the necessary infrastructure, including external communication and hardware accelerator integration. Partial dynamic reconfiguration of the FPGA is supported to further increase the efficiency of the implementation. In VEDLIoT, this will be utilized in particular to switch between different DL accelerators, e.g., to adapt to changing environmental conditions or power budgets. Therefore, accelerators with different power, performance, and accuracy footprints can be selected at runtime. The base design was created with the Xilinx Vitis Core Development Kit (2021.2) enhanced by an script environment to automate configuration and for quick adaption to other platforms. The scripts cover the hardware platform as well as the software infrastructure including the Linux environment. For a first evaluation, the base design is used in combination with the Xilinx DPU. The DPU provides a high level of flexibility, e.g., with respect to the peak number of operations per clock cycles, having a significant impact on performance, resource requirements as well as power. We have evaluated performance and energy efficiency for a wide variety of DPU configurations and devices, ranging from small (ZU3EG) to large (ZU15EG) UltraScale+ devices that can be integrated into RECS as well as for Xilinx Versal (VC1902), based on the VCK190 evaluation system. In the next Section, FPGA implementations are named by the integrated DPU, e.g., B4096 refers to a DPU which a theoretical maximum performance of 4096 operations per clock cycle. For the Versal, C32B6 comprises 6 batch handlers, each utilizing 32 AI engine cores.

III. Evaluation and benchmarking

A. Methodology

A set of common CNN models from the VEDLIoT Model Zoo [1] was used to evaluate the different accelerators and their optimizing toolchains typically provided by the hardware manufacturers. The models used in this evaluation are three state-of-the-art CNNs, namely ResNet50 [4], MobileNetV3 [5], and YoloV4 [6]. All models come from the domains of image recognition and image classification and were represented using the Open Neural Network Exchange (ONNX) [7], an open standard for ML algorithms.

Hence, for evaluation, the two mainstream benchmarking datasets *Common Objects in Context (COCO)* [8], a large-scale object detection, segmentation, and captioning database and *ImageNet* [9], the most commonly used dataset for image classification in the Large Scale Visual Recognition Challenge (ILSVRC), are used. This dataset contains 1000 object classes and contains 1,281,167 training images, 50,000 validation images, and 100,000 test images. Three versions of each model with three precision levels have been evaluated. The first version is the original trained model with 32-bit floating-point precision (FP32). The other two are quantized versions of the original model for 16-bit floating-point (FP16) and 8-bit integer precision (INT8). Table I summarizes the toolchains used for evaluation.

The evaluation metrics are driven by the requirements of the use cases targeted in the context of this project. The set of evaluated metrics is divided into four different categories: system metrics (platform and I/O); performance metrics (applications); quality metrics; and efficiency metrics. Under the system metrics category, we evaluated Peak performance (in GOPS), and Idle Power (in W). Under the performance metrics category, we evaluated the inference time, the achieved performance, the memory utilization, and energy per inference. Under the quality metrics category, we evaluated the models' accuracy and Mean Average Precision (mAP). And under the efficiency metrics, we evaluated the power efficiency in GOPS/W.

TABLE I: Toolchains used for evaluation

Hardware	Toolchain	Version
Nvidia	TensorRT SDK	7.1.3, and 8.0.1 [10]
Intel	OpenVINO	2021.4.1 [11]
Xilinx	Vitis AI	1.3 and 2.5 (Versal)
	Vitis	2021.2 and 2022.1 (Versal)
Google Coral TPU	TensorFlow [12]	2.4 and 2.5
	TensorFlow Lite	2.4 and 2.5
Hailo-8	Hailo Software Suite	4.8.1

Two quality metrics have been evaluated, each suiting the domain of the targeted CNN. For image classification, the most significant quality metric is the accuracy, representing the number of correct classifications over the number of images. The accuracy was measured in two ways: *top-1* accuracy measures how often the model prediction with the highest probability matches the ground truth. *top-5* accuracy describes if the top 5 highest-probability predicted answers include the ground truth. For object detection, the corresponding performance metric is the Mean Average Precision (mAP or mAP@X). mAP@X is the area under the precision-recall curve given an Intersection over Union (IoU) threshold X. For example, an mAP(.50) measurement means that a positive detection needs to have a minimum intersection over union (IoU) of 50 %, and that everything below will be marked as false detection with a precision of 0 %. Another form of mAP is mAP@X:Y, which is calculated as the average AP over a range of minimum IoUs, we reported the mAP@X:Y from X= 0.5 and Y= 0.95, with a step size of 0.05.

To measure power consumption, we used the utilities provided by the hardware vendors. When these were not available, the power consumption of the hardware was determined using lab instruments. For NVIDIA accelerators, the two utilities Tegrastats and nvidia-smi were used. An exception was the Jetson-Nano, where due to the lack integrated tools the external tool *Watts up PRO* was used. For Intel Myriad and its host module, a Tektronix MDO4054B oscilloscope was employed. The power consumption of the Google Coral TPU and its host module was measured using the same oscilloscope. For Hailo-8, the power measurements were done inside the RECS testbed. The Hailo-8 was plugged into the PCIe-port of Intel Xeon-D1577 and power was measured individually excluding the power consumed by the CPU module. For the FPGA systems, the complete system power has been measured, including external memory and I/O interfaces.

It should be noted that due to space limitation, in this paper only the evaluation results are presented for the YoloV4 model. This is without loss of generality as the conclusions in this paper apply also to the results for the rest of the tested models. The full evaluation results are available in [3].

B. Evaluation

As described in III-A, the software toolchains for the accelerators under test are vendor dependent and differ for most of the architectures. Although the DL models come from the same source, we needed to make sure that all devices perform the same computations and generate comparable results. Therefore, the mAP(.50) and mAP(.50:.95) were measured for every device to validate that the toolchains of the different vendors deliver comparable results. One finding of this paper is that the mAP is mainly depending on the software toolchain used to compile the model and on the quantization that has great influence too. As a result, the mAP was grouped into vendor and quantization (FP32, INT8) classes, as shown in figure 3.

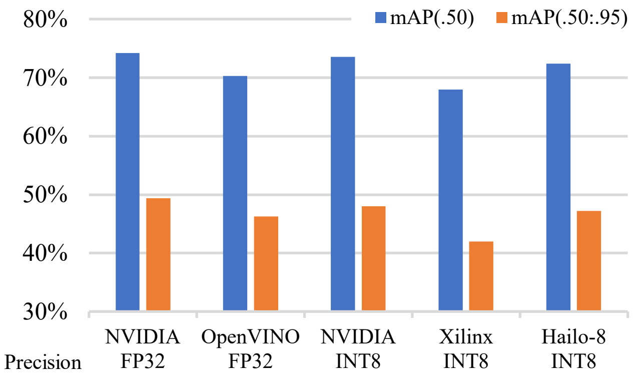


Fig. 3: Accuracy evaluation of YoloV4

The NVIDIA FP32 class combines all results for NVIDIA devices using 32-bit floating point (FP32) quantization. The OpenVINO FP32 class combines the x86-based processors and the Myriad DL accelerator using FP32 quantization.

In addition, all tests were performed using FP16 quantization, since it shows minor deviations from FP32 (<0.1 %), only FP32 and INT8 are depicted here. For the NVIDIA INT8 class that combines all NVIDIA devices using 8-bit integer quantization, the quantization needed to be done manually providing training data from the COCO dataset to the toolchain. The Xilinx INT8 and Hailo-8 INT8 classes were measured using pre-quantized models from each vendors modelzoo. Own attempts to quantize the YoloV4 model to these classes resulted in poor precision results. Quirks in the specific toolchains and hardware seem to have a significant impact on quantization, directly showing the impact on precision.

Figure 3 compares all mAP for all architectures that have been tested with the YoloV4 model. Most of the architectures show slight deviations of <5 %, while the Xilinx INT8 result is nearly 8 % lower. Further investigations lead to the analysis of Recall-Precision gradients for each of the 80 classes YoloV4 is trained on. An example for the analysis of two classes is shown in figure 4. It shows the mAP(.50) Recall-Precision gradients, where detected objects with an IoU larger than 50 % are positive detections and are displayed with their corresponding precision. All objects with an IoU smaller than 50 % are negative detections that are set to a precision of 0 %, explaining why orange and yellow precisions aren't used in the figure. Class I (toothbrush) is the class showing the

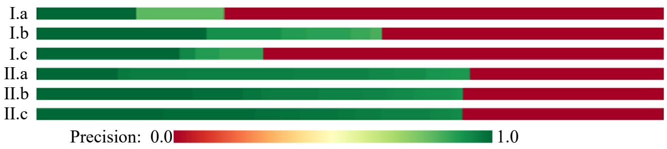


Fig. 4: mAP(.50) Recall-Precision gradients using INT8 for classes I: toothbrush II: vase and accelerators a: NVIDIA b: Hailo-8 c: Xilinx

highest deviation for INT8 quantization between the tested devices. The accelerators (a: NVIDIA) and (c: Xilinx) perform relatively poor compared to (b: Hailo-8). This is by far the class with the highest deviation compared to e.g., class II (vase) where all of the accelerators perform nearly the same. In VEDLIoT Deliverable D3.3 [3] the detailed results of this analysis are shown, comparing all 80 COCO classes per accelerator for floating point and integer quantization. Since most of the classes behave like class II we are very confident that the accelerators in our evaluation are computing the same tasks and that our results are comparable.

The evaluation in figure 5 is an overview of YoloV4 performance results delivered in D3.3 [3], it shows the performance in GOPS over the power in Watt. In D3.3 similar graphs for ResNet50 and MobileNetV3 are shown, due to the length of this publication only YoloV4 results are shown here. The small notations next to the accelerators (B1, B4, B8) refer to batch sizes 1, 4 and 8. It needs to be mentioned that all PCIe-based accelerators (Myriad, GTX1660, V100, A100, Hailo-8) have been power measured without the host system. For Hailo-8, an additional measurement including the Xavier NX as host system (Hailo-8 + Host) has been performed. In this case the Xavier NX was measured without the GPU to show the efficiency of Hailo-8 in a running system. Combining the respective power and performance values represents energy efficiency in GOPS/W which is also visualized in figure 6.

With figure 5 VEDLIoT provides a starting point when it comes to hardware decision for a wide range of use cases. Power domains, as highlighted in figure 2, can be directly transferred to choose best suited DL accelerator for given performance requirements and power budget.

Two x86 systems (D1577, Epyc3451) have been measured as basis, in order to show advances of DL accelerators over classical processing systems. When it comes to energy efficiency, interesting platforms for the VEDLIoT project are the Hailo-8, Xavier NX, Xavier AGX, VC1902, Orin AGX and A100, serving in different performance domains from low-power/embedded (Hailo-8, Xavier NX) over edge computing (Xavier AGX, VC1902, Orin AGX) to high performance computing (A100). The Hailo-8 and Xavier NX are perfectly suited for Far Edge Computing as depicted in figure 1. The u.RECS, newly designed in VEDLIoT, can be equipped with one or two of these computing devices providing maximum efficiency for low power budgets, e.g the VEDLIoT automotive use case [1]. The Xavier AGX, VC1902 and Orin AGX fit into the Near Edge Computing platform t.RECS, supporting up to three of these modules. One example for this is the VEDLIoTs smart home use case, utilizing multiple Jetson AGX-based microservers in parallel [1]. The A100 can be populated into the RECS|Box Cloud Computing platform. Within VEDLIoT it serves for model training and validation. But also the older and less efficient accelerators (Nano, Myriad, TX2) have their right to exist. Even so they lack performance, their price-tag is very interesting for some use-cases, e.g., some of the VEDLIoT open call projects [1] decided to use them.

In addition to the ASIC- and GPU-based accelerators, three reconfigurable devices (ZU3, ZU15, VC1902) have been analyzed here. The Xilinx Zynq devices (ZU3, ZU15) on the one hand show relatively low performance compared to the dedicated accelerators, due to the fact that they are basic FPGAs that implement the Xilinx DPU accelerator in their fabric. The Xilinx Versal (VC1902) on the other hand benefits

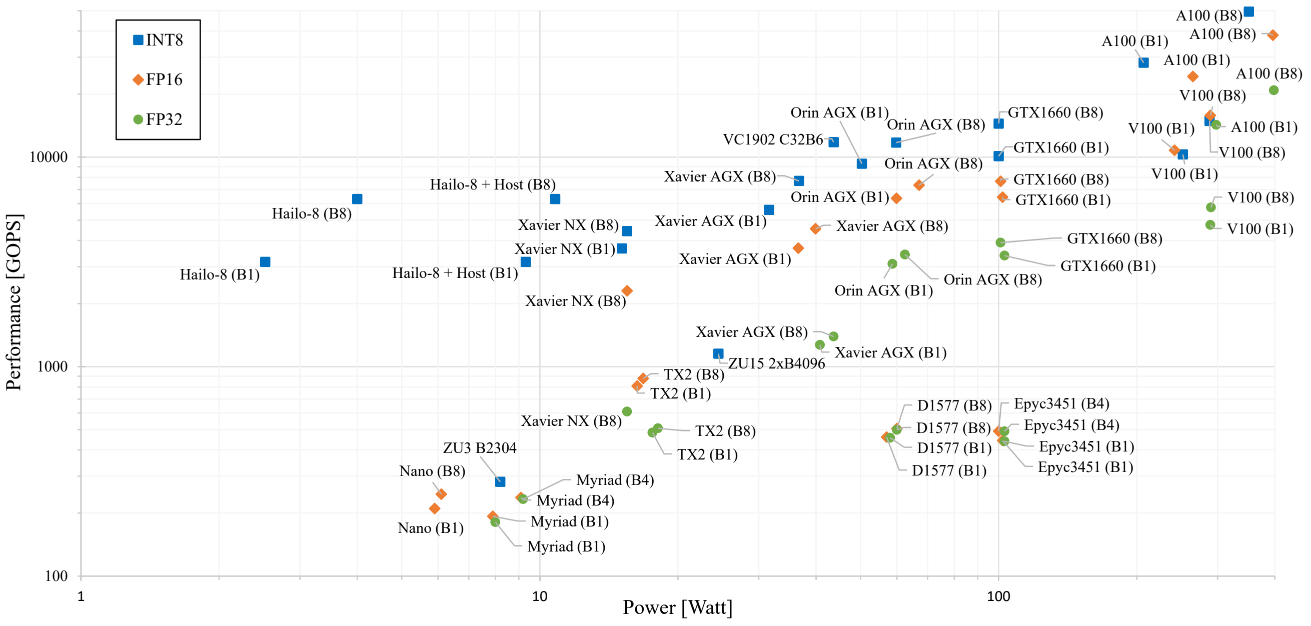


Fig. 5: Performance evaluation of YoloV4

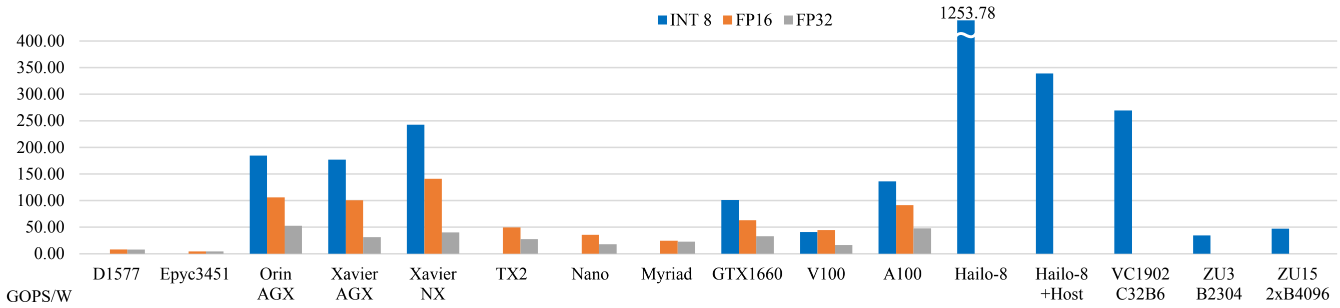


Fig. 6: Efficiency evaluation of YoloV4

from its integrated DL accelerators providing significantly higher performance and energy efficiency. Compared to all evaluated devices it shows the best energy efficiency using INT8 quantization.

Comparing energy efficiency in figure 6 it can be clearly seen, that classical processing systems (D1577, Epyc3541) lack behind. Even older DL accelerator (TX2, Nano, Myriad) provide higher efficiency. Newer GPU-based accelerators (Xavier NX, Xavier AGX, Orin AGX) provide very good efficiency, they are outperformed only by ASIC-based accelerators (Hailo-8, VC1902). Keeping in mind that all PCIe-based accelerators are power measured without their host system, the big lead of the Hailo-8 without host system is put into perspective compared to Xavier NX and VC1902 that operate stand-alone.

IV. Summary

The main topic of this paper is the evaluation of hardware, in particular accelerators, for Deep Learning applications. The RECS hardware platforms were introduced supporting cloud computing to near edge computing appliances. The new far edge u.RECS server expands these platforms towards IoT scenarios. Especially for IoT scenarios with low power budgets the energy efficiency is crucial, which is only achieved by using specialized hardware accelerators. A set of relevant accelerators was presented and classified into three different performance groups according to their processing capabilities. Besides ASIC- and GPU-based accelerators, the VEDLIoT project has an emphasis on reconfigurable architectures, presenting a DPU-based FPGA architecture for easy integration of dedicated DL algorithms.

The methodology of the evaluation was described in detail, discussing the used DL models, corresponding datasets and used specific toolchains. The performance and efficiency metrics GOPS and GOPS/W were introduced as well as the quality metrics mAP(0.50) and mAP(0.50:0.95) used for YoloV4. The power measurement used for this evaluation was described.

Since toolchains are vendor specific, an evaluation of the accuracy, of the model running on different architectures, was performed. In depth analysis of Recall-Precision gradients per class show that results of different architectures using different toolchains are still comparable. The YoloV4 evaluation shows an extensive overview of modern DL accelerators and their performance as well as their energy efficiency. The outcome of this paper provides a guideline for hardware selection in the area of DL accelerator, ranging from Far Edge Computing up to Cloud Computing.

References

[1] Martin Kaiser, Rene Griessl, Nils Kucza, et al. VEDLIoT: Very Efficient Deep Learning in IoT. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 963–968, 2022.

[2] Rene Griessl, Karol Gugala, Elaheh Malekzadeh, et al. D 3.1 – Evaluation of existing architectures and compilers for DL, October 2021. VEDLIoT project deliverable.

[3] Rene Griessl, Marco Tassemeiner, Pedro Trancoso, Karol Gugala, et al. D 3.3 – Evaluation of the DL accelerator designs, October 2022. VEDLIoT project deliverable.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[6] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[7] Junjie Bai, Fang Lu, Ke Zhang, et al. ONNX: Open Neural Network Exchange. https://github.com/onnx/onnx, 2019.

[8] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312, 2014.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Huang Rao, Chen et al. Tensorrt. https://github.com/NVIDIA/TensorRT, 2013.

[11] Paramuzov Lavrenov, Churaev et al. OpenVINO. https://github.com/openvinotoolkit/openvino, 2013.

[12] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.