Review

Survey of Deep Learning Accelerators for Edge and Emerging Computing

Md Shahanur Alam * , Chris Yakopcic , Qing Wu , Mark Barnell , Simon Khan , Tarek M. Taha

Posted Date: 10 July 2024

doi: 10.20944/preprints202407.0025.v2

Keywords: AI Accelerator; AI Frameworks; Deep Learning; Edge Computing; Low Power Applications; Quantization; PIM or CIM Computing; Neuromorphic Computing

Review

Survey of Deep Learning Accelerators for Edge and Emerging Computing

Shahanur Alam [1], Chris Yakopcic [1], Qing Wu [2], Mark Barnell [2], Simon Khan [2] and Tarek M. Taha [1]

[1] Department of Electrical and Computer Engineering, University of Dayton, Dayton, OH 45469, USA
[2] Information Directorate, Air Force Research Laboratory, Rome, NY, USA
* Correspondence: alamm8@udayton.edu

Abstract: The unprecedented progress in Artificial Intelligence (AI), particularly in deep learning algorithms with ubiquitous internet connected smart devices, has created a high demand for AI computing on the edge devices. This review studied commercially available edge processors, and the processors that are still in industrial research stages. We categorized state-of-the-art edge processors based on the underlying architecture, such as dataflow, neuromorphic, and Processing in-Memory (PIM) architecture. The processors are analyzed based on their performance, chip area, energy efficiency, and application domains. The supported programming frameworks, model compression, data precision, and the CMOS fabrication process technology are discussed. Currently, most of the commercial edge processors utilize dataflow architectures. However, emerging non-von Neumann computing architectures have attracted the industry in recent years. Neuromorphic processors are highly efficient for performing computation with fewer synaptic operations, and several neuromorphic processors offer online training for secured and personalized AI applications. This review found that the PIM processors show significant energy efficiency and consume less power compared to dataflow and neuromorphic processors. The future direction of the industry would be to implement state-of-the-art deep learning algorithms in emerging non-Von Neumann computing paradigms for low power computing on the edge devices.

Keywords: AI accelerator; AI frameworks; deep learning; edge computing; low power applications; quantization; PIM or CIM computing; neuromorphic computing

1. Introduction

Artificial intelligence, and in particular deep learning, is becoming increasingly popular in edge devices and systems. Deep learning algorithms require significant amounts of computations ranging from a few million to billions of operations based on the depth of the Deep Neural Network (DNN) models, and thus, there is an urgent need to process these efficiently. As shown in Figure 1, two possible approaches for processing deep learning inference on edge devices are directly on the device using highly efficient processors, fog, or cloud computing. A key benefit of fog/cloud-based processing is that large, complex models can be run without overburdening the edge device. The drawbacks of this approach are the need for a reliable communications channel, communications cost, communications delay, and potential loss of privacy.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0025.v2

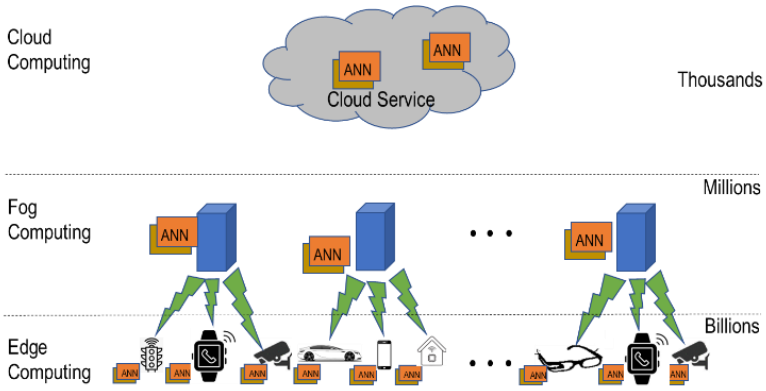

Figure 1. Illustration of edge computing with cloud interconnection.

In situations where a rapid response is needed, privacy is paramount, or a reliable communications channel may not always be available, processing of the deep learning network on the edge device or system is the only option [1–3]. As a result, a large amount of academic and industrial research is being done to develop efficient deep learning edge processors [3]. Several companies have already announced or have started selling such processors. This paper provides details on these commercial deep learning edge processors and compares their performances based on manufacturer provided information. Additionally, the paper delves into the frameworks and applications related to these processors. The scope of edge computing includes end devices and edge nodes [4]. End devices include smartphones, wearables, autonomous cars, gadgets, and many more. Edge nodes are switches, routers, micro data centers, and servers deployed at the edge [5,6]. Table 1 lists some of the key characteristics of edge deep learning processors that are considered in this paper.

There are multiple types of AI accelerators enabling DNN computing: Central Processing Unit (CPU), Graphics Processing Unit (GPU), Tensor Processing Unit (TPU), Application Specific Integrated Circuit (ASIC), System on-Chip (SoC), Processing in-Memory (PIM), and Neuromorphic processor. ASIC, SoC, TPU, PIM, and Neuromorphic systems are mainly targeted for low-power AI applications in edge and IoT devices. Google introduced different versions of the TPU that are used in the Google cloud and in the edge for training and inference [7]. Neuromorphic processors are non-Von Neumann computing systems that mimic human cognitive information processing systems. They generally utilize spiking neural networks (SNN) for processing [8,15]. Several tech companies, including Intel and IBM [8–10] have developed brain-inspired neuromorphic processors for edge applications. PIM is another non-Von Neumann computing paradigm that eliminates the data transfer bottleneck by having the computation take place inside a memory array in a highly parallel fashion [16–20].

Table 1. Brief Scopes of The Paper.

Architecture	Precision	Process (nm)	Metrics	Frameworks	Algorithm/Models	Applications
GPU	FP-8,16,32	4	Area	Tensorflow (TF)	SNN	Defense
TPU	BF-16	5	Power	TF Lite	MLP	Healthcare
Neuromorphic	INT-1,2,4,8,16	7	Throughput	Caffe2	CNN	Cyber Security
PIM		10	Energy	Pytorch	VGG	Vehicle
SoC		14	Efficiency	MXNet	ResNet	Smartphone
ASIC		16		ONNX	YOLO	Transportation
		20		MetaTF	Inception	Robotics
		22		Lava	MobileNet	Education
		28		Nengo	RNN	UAV Drones

 doi:10.20944/preprints202407.0025.v2

		40		OpenCV DarkNet	GRU BERT LSTM	Communication Industry Traffic Control

PIM technology reduces the data movement and latency compared to traditional architectures and makes the computations significantly more efficient. Edge processors usually perform inference with highly optimized DNN models. The models are often compressed to reduce the number of computations, and the weight precision is usually quantized from the floating-point (FP) format normally used in training. The quantized integer (INT) and brain-float (BF) are used in inference processors. Typically, INT4, INT8, INT16, FP16, or BF16 numerical precision is used in the inference processor. However, recently released processors from multiple startups can compute with very low precision while trading off accuracy to some extent [21].

The current trend of computing technology is to enable data movement faster for higher speed and more efficient computing. To achieve this, AI edge processors need some essential prerequisites: lower energy consumption, smaller area, and higher performance. Neuromorphic and PIM processors are becoming more popular for their higher energy efficiency and lower latency [9,10,19,20]. However, a single edge processor usually does not support all types of DNN networks and frameworks. There are multiple types of DNN models, and each usually excels at particular application domains. For example, Recurrent Neural Networks (RNN), Long-Short-Term-Memory (LSTM), and Gated Recurrent Unit (GRU) are suitable for natural language processing [22–28], but Convolutional Neural Networks (CNN), Residual Neural Network (ResNet), and Visual Geometry Group (VGG) networks are better for detection and classification [29–31].

The CMOS technology node used for fabricating each device has a significant impact on its area, energy consumption, and speed. TSMC currently uses 3nm extreme Ultraviolet (UV) technology for the Apple A17 processor [32]. TSMC is currently aspiring to develop 2nm technology by 2025 for higher performance and highly energy-efficient AI computing processors [33]. Samsung's smartphone processor Exynos 2200 is in the market developed with 4nm technology [34]. Intel utilized its Intel-4/7nm technology for its Loihi 2 neuromorphic processor [9].

This article provides a comprehensive review of commercial deep learning edge processors. Over 100 edge processors are listed along with their key specifications. We believe this is the most comprehensive technical analysis at present. The main contributions of this review are:

1. It provides a comprehensive and easy to follow description of the state-of-the-art edge devices and their underlying architecture.
2. It reviews the supported programming frameworks of the processors and general model compression techniques to enable edge computing.
3. The study has analyzed the technical details of the processors for edge computing and provides charts on hardware parameters.

This paper is arranged as follows: section 2 describes key deep learning algorithms very briefly. Section 3 describes model compression techniques commonly used to optimize deep learning networks for edge applications. Section 4 discusses the frameworks available for deep learning AI applications. Section 5 describes the frameworks for developing AI applications on SNN processors. The processors are reviewed briefly in section 6. Section 7 discusses the data on the processors and performs a comparative analysis. A brief summary of this review study is presented in section 8.

 doi:10.20944/preprints202407.0025.v2

2. Deep Learning Algorithms in Edge Application

Deep learning (DL) is a subset of AI and machine learning. It consists of multilayered artificial neural network architectures that optimize the network learning parameters to recognize the patterns and sequences for numerous applications. The networks can be trained for specific tasks, such as speech recognition [35], image recognition [36,37], security [38], anomaly detection [39], and fault detection [40]. Deep learning algorithms can be classified into the following categories: supervised, semi-supervised, unsupervised, and deep reinforcement learning [41,42].

This study is focused on AI accelerators for edge/IoT applications. Supervised and semi-supervised DL categories are usually trained on high-performance computing systems and then deployed to edge devices. Supervised learning models utilize labeled data samples. These models usually extract key features from incoming data samples and use the features to classify the sample. One of the most popular categories of supervised DL networks is CNNs [42]. Some common CNN architectures include VGG [43], ResNet [44], and GoogleNet [45]. Semi-supervised neural networks use a few labels to learn the categories and could be generative models or time-based sequence learning models. The semi-supervised topologies include GAN, GRU, RNN, and LSTM. The internal layers of these NN models are composed of CNN and fully connected network topologies. A number of edge processors support the semi-supervised network models for automation applications. For example, DeepVision (now Kinara) introduced ARA-1 (2020) and ARA-2 (2022) [46], which target autonomous applications, such as robotics, autonomous vehicles, smart tracking, and autonomous security systems. Kneron introduced KL720 in 2021, which supports semi-supervised network topologies for a wide range of applications [47]. In 2021, Syntiant released a new PIM AI processor for extreme edge applications, accommodating supervised and semi-supervised network topologies and supporting CNN, GRU, RNN, and LSTM topologies [20].

The computational complexity of DL models is a barrier to implementing these models for resource constrained edge or IoT devices. For edge applications, the deep neural network should be designed in an optimized way that is equally efficient without losing accuracy significantly. Common deep learning application areas in the edge include [48–55]: image classification, object detection, object tracking, speech recognition, health care, and natural language processing (NLP). This section will discuss some lightweight DL models for edge applications to perform classification and object detection.

i. Classification

Classification is probably the most popular use of CNNs and is one of the key applications in the computer vision field [56–58]. While larger networks with higher accuracies are utilized in desktop and server systems, smaller and more highly efficient networks are typically used for edge applications.

SqueezeNet [59,60] utilizes a modified convolutional model that is split into squeeze and expand layers. Instead of 3x3 convolution operations seen in typical CNNs, a much simpler 1x1 convolution operation is used. SqueezeNet achieves AlexNet levels of accuracy with 50x fewer network parameters [60]. Using model compression techniques, SqueezeNet can be compressed to 0.5 MB, which is about 510x smaller than AlexNet.

MobileNet [61] was created by Google and is one of the most popular DL models for edge applications. MobileNet substitutes the traditional convolution operation with a more flexible and efficient depthwise separable operation, significantly reducing computational costs. The depthwise separable technique performs two operations: depthwise convolution and pointwise convolution. There are three available versions of MobileNet networks: MobileNet v1 [61], MobileNet v2 [62], and MobileNet v3 [63]. MobileNet v2 builds on MobileNet v1 by adding a linear bottleneck and an inverted residual block at the end. The latest MobileNet v3 utilizes NAS (Neural Architecture Search) and NetAdapt to design a more accurate and efficient network architecture for inference applications [63].

 doi:10.20944/preprints202407.0025.v2

ShuffleNet [64] utilizes group convolution and channel shuffle to reduce computation complexity. It increases accuracy by retraining with minimal computational power. There are two versions of ShuffleNet, ShuffleNet v1 and ShuffleNet v2 [64,65].

EfficientNet is a family of the convolutional network model scaled from other models. It can uniformly scale all the network dimensions, such as width, depth, and resolution by using a compound coefficient [66]. The scaling method facilitates the development of a family of networks. Unlike other DL models, the EfficientNet model focuses not only on accuracy but also on the efficiency of the model.

ii. Ditection

Object detection is an important task in computer vision that identifies and localizes all the objects in an image. This application has a wide range of applications, including autonomous vehicles, smart cities, target tracking, and security systems [67]. The broad range of object detection and DL network applications are discussed in [68,69]. DL networks for object detection can be categorized into two types: i) single-stage (such as SSD, YOLO, and CenterNet) and ii) two-stage (such as Fast/Faster RCNN). There are multiple criteria for choosing the right architecture for the edge application. Single-stage detectors are computationally more efficient than two-stage architecture, making them a better choice for edge applications. For example, YOLO v5 demonstrates better performance compared to Faster-RCNN-ResNet-50 [67].

iii. Speech Recognition and Natural Language Processing

Speech recognition and natural language processing are becoming increasingly important applications of deep learning. Speech emotion and speech keyword recognition are the objectives of speech recognition. The process includes multiple state-of-the-art research fields, such as AI, pattern recognition, signal processing, and information theory. Apple's Siri and Google's Alexa illustrate the potential applications of speech recognition and manifest better computer-human interfacing. RNN based neural networks and time delay DNN (TDNN) are popular choices for speech recognition [70]. Combined networks, such as TDNN-LSTM [71] or RNN-LSTM, are also popular choices for speech recognition [72].

Detailed analysis of deep neural networks for NLP can be found in [73,74]. Important applications of NLP are machine translation, named entity recognition, question answering system, sentiment analysis, spam detection, and image captioning. An early NLP model was sequence2sequence learning, based on RNNs. More recently, NLP was boosted by the advent of the transformer model, BERT [75]. BERT utilized an attention mechanism that learned contextual relations between words [75]. Other state-of-the-art NLP models are GPT-2 [76], GPT-3 [77], GPT-4 [78], and switch transformer [79]. However, these models run on HPC systems and are thus not compatible with edge devices. DeFormer [80], MobileBERT [81], and EdgeBERT [82] are some of the examples of NLP models targeted for edge devices. A more detailed discussion on NLP models for edge devices can be found in [83].

Syntiant [20] has recently been building tiny AI chips for voice and speech recognition and has attracted attention in the tech industry. Syntiant's Neural Decision Processors (NDPs) are certified by amazon for use in Alexa-based devices [84]. Other voice recognition AI chips include NXP's i.MX8, i.MX9x [85–87] and M1076 from Mythic [88–90]. LightSpear 2803S from Gyrfalcon can be utilized for NLP [91,92]. IBM unveiled its NorthPole edge processor for NLP applications at the HotChips 2023 conference [299].

3. Model Compression

Unoptimized DL models contain considerable redundancy in parameters and are generally designed without consideration of power or latency. Lightweight and optimized DL models enable AI application on edge devices. Designing effective models for running on resource-constrained systems is challenging. DNN model compression techniques are utilized to convert unoptimized

 doi:10.20944/preprints202407.0025.v2

models to forms that are suitable for edge devices. Model compression techniques are studied extensively and discussed in [93–98]. The techniques include parameter pruning, quantization, low-rank factorization, compact filtering, and knowledge distillation. In this section, we will discuss some of the key model compression techniques.

i. Quantization

Quantization is a promising approach to optimize the DNN models for edge devices. Data quantization for edge AI has been studied extensively in [94–101]. Parameter quantization takes a DL model and compresses its parameters by changing the floating-point weights to a lower precision to avoid costly floating-point computations. As shown in Table 2, most edge inference engines support INT4, 8, or 16 precisions. Quantization techniques can be taken to the limit by developing Binary Neural Networks (BNN) [101]. The BNN uses a single bit to represent activations and reduces memory requirements. Leapmind is the pioneer of low precision computations in their edge processor, Efficiera [21]. It is an ultra-low power edge processor and can perform AI computations with 1 bit weights and 2 bit activations.

Recent hardware studies show that lower precision does not have a major impact on inference accuracy. For example, Intel and Tsinghua University have presented QNAP [102], where they utilize 8 bits for weights and activations. They show an inference accuracy loss of only 0.11% and 0.40% for VGG-Net and GoogleNet, respectively, when compared to a software baseline with the ImageNet dataset. Samsung and Arizona State University have experimented with extremely low precision inference in PIMCA [103], where they utilize 1 bit for weights and activations. They show that VGG-9 and ResNet-18 had accuracy losses of 3.89% and 6.02% respectively.

Lower precision increases the energy and area efficiency of a system. PIMCA can compute 136 and 35 TOPS/W in 1 and 2 bit precision, respectively for ResNet-18. TSMC [104] has studied the impact of low precision computations on area efficiency. They show 221 and 55 TOPS/mm^2 area efficiency in 4- and 8-bit precision. Thus, with 4-bit computation, they achieve about 3.5x higher computation throughput per unit area compared to 8-bit computation.

Brain-Float-16 (or BF-16) [105] is a limited precision floating point format that is becoming popular for AI applications in edge devices. BF16 combines certain components of FP32 and FP16. From FP16, the BF16 utilizes 16 bits overall. From FP32, BF16 utilizes 8 bits for the exponent field (instead of 5 bits for FP16). A key benefit of BF16 is the format gets the same dynamic range and inference accuracy as of FP32 [106]. BF16 speeds up the MAC operation in edge devices to enable faster AI inference on the edge devices. Both the GDDR6-AiM from SK Hynix [107] and Pathfinder-1600 from Blaize [108,109] support BF16 for AI applications. The supported precision levels of various edge processors are presented in Table 2.

ii. Pruning

Pruning is the technique to remove unnecessary network connections to make the network lightweight for deploying on edge processors. Several studies [94–101,110–112] show that upto 91% of weights in AlexNet can be pruned with minimal accuracy reduction. Various training methods have been proposed to apply pruning to pre-trained networks [101]. Pruning however has drawbacks, such as creating sparsity in the weight matrices. This sparsity leads to unbalanced parallelism in the computation and irregular access to the on-chip memory. Several techniques have been developed [113,114] to reduce the sparsity.

iii. Knowledge Distillation

Knowledge distillation, introduced by B. Christian et al. [115], is a technique where the knowledge of an ensemble of larger networks is transferred to a smaller network without loss of validity. This can reduce the computational load significantly. The effectiveness of knowledge distillation is studied extensively in [94–101,116–120], where the authors show that the distillation of knowledge from a larger regularized model into a smaller model works effectively. Various

 doi:10.20944/preprints202407.0025.v2

algorithms have been proposed to improve the process of transferring knowledge, such as adversarial distillation, multi-teacher distillation, cross-modal distillation, attention-based distillation, quantized distillation, and NAS based distillation [121]. Although knowledge distillation techniques are mainly used for classification applications, they are also applied to other applications, such as object detection, semantic segmentation, language modeling, and image synthesis [122].

4. Framework for Deep Learning Networks

At present, the majority of edge AI processors are designed for inference only. Network training is typically carried out on higher performance desktop or server systems. There are a large variety of software frameworks to train deep networks and also to convert them into lightweight, suitable for edge devices. Popular DNN frameworks include Tensorflow (TF) [123], Tensorflow Lite (TFL) [124], PyTorch [125], PyTorch mobile [126], Keras [127], Caffe2 [128], OpenCV [129], ONNX [130], and MXNet [131]. Some of these frameworks support a broad class of devices, such as android, iOS, or Linux systems.

TFL was developed by Google and supports interfacing with many programming languages (such as Java, C++, Python). It can take a trained model from TensorFlow and apply model compression to reduce the amount of computations needed for inference.

ONNX was developed by the PyTorch team to represent traditional machine learning and state-of-the-art deep learning models [130]. The framework is interoperable across popular development tools, such as PyTorch, Caffe2, and Apache MXNet. Many of the current AI processors support the ONNX framework, such as Qualcomm SNPE, AMD, ARM, and Intel [132].

PyTorch mobile was developed by Facebook and allows a developer to train AI models for edge applications. The framework provides a node-to-node workflow that enables the clients to have a privacy-preserving learning environment via collaborative or federated learning [125,126]. It supports XNNPACK floating point kernel libraries for ARM CPUs and integrates QNNPACK for quantized INT8 kernels [126].

Caffe2 is a lightweight framework developed by Facebook [128]. This framework supports C++ and Python APIs, which are interchangeable and helps to develop prototypes quickly that could potentially be optimized later. Caffe2 integrates with Android Studio and Microsoft Visual Studio for mobile development [128]. Caffe2Go is developed to embed in mobile apps for applying a full-fledged deep learning framework for real-time capture, analysis, and decision making without the help of a remote server [133].

Facebook uses Pytorch Mobile, Caffe2 and ONNX for developing their products. Pytorch is used for the experiment and rapid development, Caffe2 is developed for aiming at the production environment, while ONNX helps to share the models between the two frameworks [130].

MXNet is a fast and scalable framework developed by the Apache Software Foundation [131]. This framework supports both training and inference with a concise API for AI applications in edge devices. MXNet supports Python, R, C++, Julia, Perl, and many other languages and can be run on any processor platform for developing AI applications [131]. As shown in Table 4, TFL, ONNX, and Caffe2 are the most widely used frameworks for AI edge applications.

Some edge processors are compatible only with their in-home frameworks. For example, Kalray's MPPA3 edge processor is compatible with KaNN (Kalray Neural Network), so any trained deep network must be converted to KaNN to run on the MPPA3 processor [13]. CEVA introduced its own software framework CEVA-DNN for converting pre-trained network models and weights from offline training frameworks (such as Caffe, TensorFlow) for inference applications on the CEVA processors [134–136]. CEVA added a retrain feature in CEVA-DNN for the Neuro-Pro processor to enable a deployed device to be updated without uploading a database to the server [134]. The developer can also use CEVA-DNN tools on a simulator or test device and then transfer the updated model to edge devices [136].

 doi:10.20944/preprints202407.0025.v2

5. Framework for Spiking Neural Networks

Spiking neural networks (SNN) utilize brain inspired computing primitives, where a neuron accumulates a potential and fires only when a threshold is crossed [137]. This means in spiking neural networks, the neurons have outputs sporadically. Thus, SNNs have much fewer neuron to neuron communications compared to deep neural networks, where all neurons always send outputs. The net result of this is that SNNs can be dramatically more power efficient than DNNs and could potentially implement a task with far fewer operations. Thus, an SNN processor with the same operations per second capability as a DNN processor could theoretically have a much higher task level throughput.

To get the highest efficiency from SNN processors, it is best to use algorithms that are developed from the ground up to use spiking neurons. Examples of such algorithms include constraint satisfaction problems [138] and genetic algorithms [139]. Several studies have examined how to implement DNNs using SNNs [140]. Davidson et al. [141] show through modeling of energies that this should not result in higher efficiency than the original DNN using the same underlying silicon technology. However, P. Blouw et al. [142] implemented keyword spotting on several hardware platforms and showed that the Loihi was about 5x more energy efficient than the Movidius deep learning processor. The remainder of this section describes some of the key frameworks for implementing SNN architectures for spiking neuromorphic processors.

Nengo is a Python based framework developed by Applied Brain Research for spiking neurons. It supports multiple types of processors, including Loihi [143] and Spinnaker [144]. Nengo is very flexible in writing code and simulating SNNs. The core framework is the Nengo ecosystem, which includes Nengo objects and NumPy based simulators. The Nengo framework has Nengo GUI for model construction and visualization tools and NengoDL for simulating deep learning models using SNNs [145].

Meta-TF [146] is a framework developed by BrainChip for edge application in the Akida neuromorphic chips [147–149]. Meta-TF takes advantage of the Python scripting language and associated tools, such as Jupyter notebook and NumPy. Meta-TF includes three Python packages [146]: 1) The Akida Python package works as an interface to the Akida neuromorphic SoC. 2) the CNN2SNN tool provides an environment to convert a trained CNN network into SNNs. Brainchip embeds the on-chip training capability in the Akida processor, and thus, the developers can train SNNs on the Akida processor directly [149]. iii) Akida Model Zoo contains pre-created network models, which are built with the Akida sequential API and the CNN2SNN tool by using quantized Keras models.

Lava is a framework currently being developed by Intel to build SNN models and map them to neuromorphic platforms [150]. The current version of the Lava framework supports the Loihi neuromorphic chips [9]. Lava includes Magma which helps to map and execute neural network models and sequential processes to neuromorphic hardware [150]. Magma also helps to estimate performance and energy consumption on the platform. Lava has additional properties, including offline training, integration with other frameworks, a Python interface, and being an open-source framework (with proper permissions). The Lava framework supports online real-time learning, where the framework adopts plasticity rules. However, the learning is constrained to access only locally available process information [150].

6. Edge Processors

At present, GPUs are the most popular platform for implementing DNNs. These, however, are usually not suitable for edge computing (except the NVIDIA Jetson systems) due to their high-power consumption. A large variety of AI hardware has been developed, many of which target edge applications. Several articles have reviewed AI hardware in broad categories, giving an overall idea of the current trend in AI accelerators [151–153]. Earlier works [2,154–156] have reviewed a small selection of older edge AI processors.

This paper presents a very broad coverage of edge AI processors and PIM processors from the industry. This includes processors already released, processors that have been announced, and processors that have been published about in research venues (such as the ISSCC and the VLSI

 doi:10.20944/preprints202407.0025.v2

conferences). This section is divided into four subsections: subsection (i) describes dataflow processors, subsection (ii) describes neuromorphic processors, and subsection (iii) describes PIM processors. All of these sections describe industrial products that have been announced or released. Finally, subsection (iv) describes the processors in industrial research.

Table 2 describes the key hardware characteristics of the commercial edge-AI and PIM-AI processors. Table 3 lists the same key characteristics for the processors from industrial research. Table 4 describes the key software/application characteristics of the processors in Tables 2 .

i. Dataflow Edge Processor

This section describes the latest dataflow processors from the industry. Dataflow processors are custom designed for neural network inference and, in some cases, training computations. The processors are listed in alphabetical order based on the manufacturer name. The data provided is from the publications or websites of the processors.

Apple released the bionic SoC A16 with an NPU unit for the iPhone 14 [157]. The A16 processor exhibits about 20% better performance with the same power consumption as their previous version, A15. It is embedded with a 6-core ARM8.6a CPU, 16-core NPU, and 8-core GPU [157]. The Apple M2 processor was released in 2022 primarily for the Macbooks, and then optimized for iPads. This processor includes a 10-core GPU and 16-core NPU [158]. M1 performs 11 TOPS with 10 W of power consumption [159]. M2 has 18% and 35% more powerful CPU and GPU for faster computations.

ARM recently announced the Ethos-N78 with an 8-core NPU for automotive applications [160]. Ethos-N78 is an upgraded version of Ethos-N77. Both NPUs support INT8 and INT16 precision. Ethos-N78 performs more than 2x better than the earlier version. The most significant improvement of Ethos-N78 is enabling a new data compression method that reduces the bandwidth and improves performance and energy efficiency [161].

Blaize released its Pathfinder P1600, El Cano AI inference processor. This processor integrates 16 graph streaming processors (GSP) that deliver 16 TOPS at its peak performance [162]. It uses a dual Cortex-A53 for running the operating system at up to 1GHz. Blaize GSP processors integrate data pipelining and support up to INT-64 and FP-8-bit operations [163].

AIMotive [164] introduced the inference edge processor Apache5, which supports a wide range of DNN models. The system has an aiWare3p NPU with an energy efficiency of 2 TOPS/W. Apache5 supports INT8 MAC and INT32 internal precision [165]. This processor is mainly targeted for autonomous vehicles [166].

CEVA [134] released the Neupro-S on-device AI processor for computer vision applications. Neupro comprises two separate cores. One is the DSP-based Vector Processor Unit (VPU), and the other is the Neupro Engine. VPU is the controller, and the Neupro Engine does most of the computing work with INT8 or INT16 precision. A single processor performs up to 12.5 TOPS, while the performance can be scaled to 100 TOPS with multicore clusters [134,135]. The deep learning edge processors are mostly employed for inference tasks. CEVA added a retraining capability to its CDNN (CEVA DNN) framework for online learning on client devices [136].

Cadence introduced the Tensilica DNA 100, which is a comprehensive SoC for domain-specific on-device AI edge accelerators [167]. It has low, mid, and high-end AI products. Tensilica DNA 100 offers 8 GOPS to 32 TOPS AI processing performance currently and predicts 100 TOPS in future releases [168]. The target application of the DNA 100 is IoTs, intelligent sensors, vision, and voice application. The mid and high-end applications include smart surveillance and autonomous vehicles, respectively.

 doi:10.20944/preprints202407.0025.v2

Table 2. Commercial Edge Processors with Operation Technology, Process Technology, and Numerical Precision.

Company	Latest Chip	Power (W)	Process (nm)	Area (mm^2)	Precision INT/FP	Performance (TOPS)	E. Eff. (TOPS/W)	Architecture	Reference
Apple	M1	10	5	119	64	11	1.1	Dataflow	[159]
Apple	A14	6	5	88	64	11	1.83	Dataflow	[242]
Apple	A15	7	5		64	15.8	2.26	Dataflow	[242]
Apple	A16	5.5	4		64	17	3	Dataflow	[157]
*AIStorm	AIStorm	0.225			8	2.5	11	Dataflow	[243]
*AlphaIC	RAP-E	3			8	30	10	Dataflow	[244]
aiCTX	Dynap-CNN	0.001	22	12	1	0.0002	0.2	Neuromorphic	[15,213]
*ARM	Ethos78	1	5		16	10	10	Dataflow	[160,161]
*AIMotive	Apache5 IEP	0.8	16	121	8	1.6-32	2	Dataflow	[164,165]
*Blaize	Pathfinder, EI Cano	6	14		64, FP-8, BF16	16	2.7	Dataflow	[162]
*Bitman	BM1880	2.5	28	93.52	8	2	0.8	Dataflow	[245,246]
*BrainChip	Akida1000	2	28	225	1,2,4	1.5	0.75	Neuromorphic	[147,148]
*Cannan	Kendrite K210	2	28		8	1.5	1.25	Dataflow	[247,248]
*CEVA	CEVA-Neuro-S		16		2, 5, 8, 12, 16	12.7		Dataflow	[134]
*CEVA	CEVA-Neuro-M	0.83	16		2, 5, 8, 12, 16	20	24	Dataflow	[135]
*Cadence	DNA100	0.85	16		16	4.6	3	Dataflow	[167,168]
*Deepvision	ARA-1	1.7	28		8,16	4	2.35	Dataflow	[169]
*Deepvision	ARA-2		16					Dataflow	[170]
*Eta	ECM3532	0.01	55	25	8	0.001	0.1	Dataflow	[249]
*FlexLogic	InferX X1	13.5	7	54	8	7.65	0.57	Dataflow	[250]
*Google	Edge TPU	2	28	96	8, BF16	4	2	Dataflow	[176,177]
*Gyrfalcon	LightSpeer 2803S	0.7	28	81	8	16.8	24	PIM	[224]
*Gyrfalcon	LightSpeer 5801	0.224	28	36	8	2.8	12.6	PIM	[224]
*Gyrfalcon	Janux GS31	650/900	28	10457.5	8	2150	3.30	PIM	[225]
*GreenWaves	GAP9	0.05	22	12.25	FP-(8,16,32)	0.05	1	Dataflow	[180,181]
*Horizon	Journey 3	2.5	16		8	5	2	Dataflow	[171]
*Horizon	Journey5/5P	30	16		8	128	4.8	Dataflow	[172,173]
*Hailo	Hailo 8 M2	2.5	28	225	4,8,16	26	2.8	Dataflow	[174,175]

 doi:10.20944/preprints202407.0025.v2

Intel	Loihi 2	0.1	7	31	8	0.3	3	Neuromorphic	[9]
Intel	Loihi	0.11	14	60	1-9	0.03	0.3	Neuromorphic	[9,218]
*Intel	Intel® Movidius	2	16	71.928	16	4	2	Dataflow	[186]
IBM	TrueNorth	0.065	28	430	8	0.0581	0.4	Neuroorphic	[10,218]
IBM	NorthPole	74	12	800	2,4,8	200 (INT8)	2.7	Dataflow	[299,304]
*Imagination	PowerVR Series3NX				FP-(8,16)	0.60		Dataflow	[182,183]
*Imagination	IMG 4NX MC1	0.417			4,16	12.5	30	Dataflow	[184]
*Imec	DIANA		22	10.244	2	29.5 (A), 0.14 (D)	14.4	PIM+Digital	[222,223]
*Kalray	MPPA3	15	16		8,16	255	1.67	Dataflow	[13]
*Kneron	KL720 AI	1.56	28	81	8,16	1.4	0.9	Dataflow	[191]
*Kneron	KL530	0.5			8	1	2	Dataflow	[192]
*Koniku	Konicore							Neuromorphic	[12]
*LeapMind	Efficiera	0.237	12	0.422	1,2,4,8,16,32	6.55	27.7	Datalow	[21]
Memryx	MX3	1	--	--	4,8,16 (W) BF16	5	5	Dataflow	[297]
*Mythic	M1108	4		361	8	35	8.75	PIM	[89]
*Mythic	M1076	3	40	294.5	8	25	8.34	PIM	[18,88,90]
*mobileEye	EyeQ5	10	7	45	4,8	24	2.4	Dataflow	[193–195]
*mobileEye	EyeQ6	40	7		4,8	128	3.2	Dataflow	[196]
*Mediatek	i350		14			0.45		Dataflow	[251]
*NVIDIA	Jetson Nano B01	10	20	118	FP16	1.88	0.188	Dataflow	[197]
NVIDIA	AGX Orin	60	7	--	8	275	3.33	Dataflow	[199]
*NXP	i.MX 8M+		14	196	FP16	2.3		Dataflow	[86,87]
*NXP	i.MX9	$4x10^{-6}$	12					Dataflow	[85]
*Perceive	Ergo	0.073	5	49	8	4	55	Dataflow	[252]
TSU & Polar Bear Tech	QM930	12	12	1089	4,8,16	20 (INT8)	1.67	Dataflow	[302]
Qualcomm	QCS8250		7	157.48	8	15		Dataflow	[200,201]
Qualcomm	Snapdragon 888+	5	5		FP32	32	6.4	Dataflow	[202–204]

 doi:10.20944/preprints202407.0025.v2

Qualcomm	Snapdragon 8 Gen2		4		4,8,16, FP16	51		Dataflow	[303]
*RockChip	rk3399Pro	3	28	729	8, 16	3	1	Dataflow	[253]
Rokid	Amlogic A311D		12			5		Dataflow	[254]
Samsung	Exynos 2100		5			26		Dataflow	[205,206]
Samsung	Exynos 2200		4		8,16, FP16			Dataflow	[255]
Samsung	HBM-PIM	0.9	20	46.88		1.2	1.34	PIM	[226,227]
Sima..ai	MLSoC	10	16	175.55	8	50	5	Dataflow	[300,301]
Synopsis	EV7x		16		8, 12, 16,	2.7		Dataflow	[209,210]
*Syntiant	NDP100	0.00014	40	2.52		0.000256	20	PIM	[228,229]
*Syntiant	NDP101	0.0002	40	25	1,2, 4,8	0.004	20	PIM	[228,231]
*Syntiant	NDP102	0.0001	40	4.2921	1, 2, 4, 8	0.003	20	PIM	[228,235]
*Syntiant	NDP120	0.0005	40	7.75	1,2,4,8	0.0019	3.8	PIM	[228,234]
*Syntiant	NDP200	0.001	40		1,2,4,8	0.0064	6.4	PIM	[228,232]
Think Silicon	NEMA®|pico XS	0.0003	28	0.11	FP16,32	0.0018	6	Dataflow	[256]
Tesla/Samsung	FSD Chip	36	14	260	8, FP-8	73.72	2.04	Dataflow	[211]
Videntis	TEMPO							Neuromorphic	[11]
Verisilicon	VIP9000		16		16, FP16	0.5-100		Dataflow	[207,208]
Untether	TsunAImi	400	16		8	2008	8	PIM	[236,237]
UPMEM	UPMEM-PIM	700	20		32, 64	0.149		PIM	[238–241]

*Processors are available for purchase; **Integer Precision is indicated by only precision number(s). Floating point precision is mentioned as FP in the precision column.

Deepvision has updated their edge inference coprocessor ARA-1 for applications to autonomous vehicles and smart industries [46. It includes eight compute engines with 4 TOPS and consumes 1.7-2.3 W of power [169]. The computing engine supports INT8 and INT16 precision. Deepvision has recently announced its second-generation inference engine, ARA-2, which will be released later in 2022 [170]. The newer version will support LSTM and RNN neural networks in addition to the networks supported in ARA-1.

Horizon announced its next automotive AI inference processor Journey 5/5P [171], which is the updated version of Journey 3. The mass production of Journey 5 will be starting in 2022. The processor exhibits a performance of 128 TOPS, and has a power of 30 W, giving an energy efficiency of 4.3 TOPS/W [172,173].

Hailo released its Hailo-8 M-2 SoC for various edge applications [174]. The computing engine supports INT8 and INT16 precision. This inference engine is capable of 26 TOPS and requires 2.5 W of power. The processor can be employed as a standalone or coprocessor [175].

Google introduced its Coral Edge TPU, which comprises only 29% of the floorpan of the original TPU for edge applications [176]. The Coral TPU shows high energy efficiency in DNN computations compared to the original TPUs which are used in cloud inference applications [178]. Coral Edge TPU supports INT8 precision and can perform 4 TOPS with 2 Watts of power consumption [176].

 doi:10.20944/preprints202407.0025.v2

Google released its Tensor processor for mobile applications, coming with its recent Pixel series mobile phone [179]. Tensor is an 8-core cortex CPU chipset fabricated with 5 nm process technology. The processor has a 20-core Mali-G78 MP20 GPU with 2170 GFLOPS computing speed. The processor has a built-in NPU to accelerate AI models with a performance of 5.7 TOPS. The maximum power consumption of the processor is 10W.

GreenWaves announced their edge inference chip GAP9 [180]. It is a very low-cost, low-power device that consumes 50 mW and performs 50 GOPS at its peak. GAP9 provides hearable developments through DSP, AI accelerator, and ultra-low latency audio streaming on IoT devices. GAP9 supports a wide range of computing precision, such as INT8, 16, 24, 32, and FP16, 32 [181].

IBM introduced the NorthPole [299], a non-Von Neumann deep learning inference engine, at the HotChips 2023 conference. The processor shows massive parallelism with 256 cores. Each core has 768KB of near-compute memory to store weights, activations, and programs. The total on-chip memory capacity is 192 MB. The NorthPole processor does not use off-chip memory to load weights or store intermediate values during deep learning computations. Thus, it dramatically improves latency, throughput, and energy consumption, which helps outperform existing commercial deep learning processors. The external host processor works on three commands: write tensor, run network, and read tensor. The NorthPole processor follows a set of pre-scheduled deterministic operations in the core array. It is implemented in 12nm technology and has 22 billion transistors taking up 800 mm^2 of chip area. The performance data released on the NothPole processor are computed based on frame/sec. The performance metrics of operations/sec in integer or floating point are unavailable in the public domain currently. However, the operation per cycle is available for different data precisions. In vector-matrix multiplication, 8, 4, and 2-bit can perform 2048, 4096, and 8192 operations/cycles. The FP16 can compute 256 operations/cycle (the number of cycles/s is not released at this time). NorthPole can compute 800, 400, and 200 TOPS with INT 2, 4, and 8 precision. The processor can be applied to a broad area of applications and can execute inference with a wide range of network models applied in classification, detection, segmentation, speech recognition, and transformer models in NLP.

Imagination introduced a wide range of edge processors with targeted applications in IoTs to autonomous vehicles [182]. The edge processor series is categorized as the PowerVR Series3NX and can achieve up to 160 TOPS with multicore implementations. For ultra-low power applications, one can choose PowerVR AX3125, which has a 0.6 TOPS computing performance [183]. IMG 4NX MC1 is a single-core Series 4 processor for autonomous vehicle applications and performs at 12.5 TOPS with less than 0.5 W of power consumption [184].

Intel released multiple edge AI processors such as Nirvana Spring Crest NNP-I [185] and Movidious [186]. Recently, they have announced a scaleable 4th generation Xeon processor series that can be used for desktop to extreme edge devices [187]. The power consumption for an ultra-mobile processor is around 9W while computed with INT8 precision. The development utilizes the SuperFin fabrication technology with 10nm process technology. Intel is comparing its core architecture to the Skylake processor, and it claims an efficient core achieves 40% better performance with 40% less power.

IBM developed the Artificial Intelligence Unit (AIU) based on their AI accelerator used in the 7-nanometer Telum chip that powers its z16 system [188]. AIU is a scaled version developed on a 5 nm process technology and features a 32-core design with a total of 23 billion transistors. AIU uses IBM's approximate computing frameworks where the computing executes with FP16 and FP32 precisions [189].

Leapmind has introduced the Efficiera for edge AI inference implemented in FPGA or ASIC [21]. Efficiency is for ultra-low power applications. The computations are typically performed in 8-, 16-, or 32-bit precision. However, the company claims that 1-bit weight and 2-bit activation can be achieved while still maintaining accuracy for better power and area efficiency. They show 6.55 TOPS at 800MHz clock frequency with an energy efficiency of 27.7 TOPS/W [190].

Kneron released its edge inference processor, KL 720, for various applications, such as autonomous vehicles and smart industry [191]. The KL 720 is an upgraded version of the earlier KL

 doi:10.20944/preprints202407.0025.v2

520 for similar applications. The revised version performs at 0.9 TOPS/W and shows up to 1.4 TOPS. The neural computation supports INT8 and INT16 precisions [191]. Kneron's most up-to-date heterogeneous AI chip is KL 530 [192]. It is enabled with a brand new NPU, which supports INT4 precision and offers 70% higher performance than that of INT8. The maximum power consumption of KL 530 is 500 mW and can deliver up to 1 TOPS [192].

Table 3. Processors, Supported Neural Network Models, Deep Learning Frameworks, And Application Domains.

Company	Product	Supported Neural Networks	Supported Frameworks	Application/benefits
Apple	Apple A14	DNN	TFL	iPhone12 series
Apple	Apple A15	DNN	TFL	iPhone13 series
aiCTX-Synsense	Dynap-CNN	CNN, RNN, Reservoir Computing	SNN	High-speed aircraft, IoT, security, healthcare, mobile
ARM	Ethos78	CNN and RNN	TF, TFL,Caffe2,PyTorch, MXNet, ONNX	Automotive
AIMotive	Apache5 IEP	GoogleNet, VGG16, 19, Inception-v4, v2, MobileNet v1, ResNet50, Yolo v2	Caffe2	Automotives, pedestrian detection, vehicle detection, lane detection, driver status monitoring
Blaize	EI Cano	CNN, YOLO v3	TFL	Fit for industrial, retail, smart-city, and computer-vision systems
BrainChip	Akida1000	CNN in SNN, Mobilenet	MetaTF	Online learning, data analytics, security
BrainChip	AKD500, 1500, 2000	DNN	MetaTF	Smart home, Smart health, Smart City and smart transportation
CEVA	Neuro-s	CNN, RNN	TFL	IoTs, smartphones, surveillance, automotive, robotics, medical
Cadence	Tensilica DNA100	FCC, CNN, LSTM	ONNX, Caffe2, TensorFlow	IoT, smartphones, AR/VR, smart surveillance, Autonomous vehicle
Deepvision	ARA-1	Deep Lab V3, Resnet-50,	Caffe2, TFL, MXNET, PyTorch	Smart retail, robotics, industrial automation,

 doi:10.20944/preprints202407.0025.v2

		Resnet-152, MobileNet-SSD, YOLO V3, UNET		smart cities, autonomous vehicles, and more
Deepvision	ARA-2	Model in ARA-1 and LSTM, RNN,	TFL, Pytorch	Smart retail, robotics, industrial automation, smart cities,
Eta	ECM3532	CNN, GRU, LSTM	---	Smart home, consumer products, medical, logistics, smart industry
Gyrfalcon	LightSpeer 2803S	CNN based, VGG, ResNet, MobileNet;	TFL, Caffe2	High performance audio and video processing
Gyrfalcon	LightSpeer 5801	CNN based, ResNet, MobileNet and VGG16,	TFL, PyTorch & Caffe2	Object Detection and Tracking, NLP, Visual Analysis
Gyrfalcon Edge Server	Janux GS31	VGG, REsNet, MobileNet	TFL, Caffe2, PyTorch	Smart cities, surveillance, object detection, recognition
GreenWaves	GAP9	CNN, mobileNet v1		DSP application
Horizon	Journey 3	CNN, mobilenet v2, efficient net	TFL, Pytorch, ONNX, mxnet, Caffe2	Automotive
Horizon	Journey5/5P	Resnet18, 50, MobileNet v1-v2, ShuffleNetv2, EfficientNet FasterRCNN, Yolov3	TFL, Pytorch, ONNX, mxnet, Caffe2	Automotive
Hailo	Hailo 8 M2	YOLO 3, YOLOv4, CenterPose, CenterNet, ResNet-50	ONNX, TFL	Edge vision applications
Intel	Loihi 2	SNN based NN	Lava, TFL, Pytorch	Online learning, sensing, robotics, healthcare

 doi:10.20944/preprints202407.0025.v2

Intel	Loihi	SNN based NN	Nengo	Online learning, robotics, healthcare and many more
Imagination	PowerVR Series3NX	Mobilenet v3, CNN	Caffe, TFL	smartphones, smart cameras, drones, automotives, wearables,
Imec & GF	DIANA	DNN	TFL, Pytorch	Analog computing in Edge inference
KoniKu	Konicore	synthetic Biology+Silicon	--	Chemical detection, aviation, security
Kalray	MPPA3	Deep network converted to KaNN	Kalray's KANN	Autonomous vehicles, surveillance, robotics, industry, 5G
Kneron	KL720 AI	CNN, RNN, LSTM	ONNX, TFL, Keras, Caffe2	wide applications from automotive to home appliances
Kneron	KL520	Vgg16, Resnet, GoogleNet, YOLO, Lenet, MobileNet, FCC	ONNX, TFL, Keras, Caffe2	Automotive, home, industry and so on.
LeapMind	Efficiera	CNN, YOLO v3, Mobilenet-v2, Lmnet	Blueoil, Python & C++ API	Home, Industrial machinery, surveillance camera, robots
Memryx	MX3	CNN	Pytorh, ONNX, TF, Keras	Automation, surveillance, agriculture, financial
Mythic	M1108	CNN, large complex DNN, Resnet50, YOLO v3, Body25	Pytorch, TFL, and ONNX	Machine Vision, Electronics, Smart Home, UAV/Drone, Edge Server
Mythic	M1076	CNN, Complex DNN, Resnet50, YOLO v3	Pytorch, TFL, and ONNX	Surveillance, Vision, voice, Smart Home, UAV, Edge Server
MobileEye	EyeQ5	DNN		Autonomous driving
MobileEye	EyeQ6	DNN		Autonomous driving
Mediatek	i350	DNN	TFL	Vision and voice, Biotech and Bio-metric measurements

 doi:10.20944/preprints202407.0025.v2

NXP	i.MX 8M+	DNN	TFL, Arm NN, ONNX	Edge Vision
NXP	i.MX9	CNN, Mobilenet v1	TFL, Arm NN, ONNX	Graphics, image, display, audio
NVIDIA	AGX Orin	DNN	TF, TFL, Caffe, Pytorch	Robotics, Retail, Traffic, Manufacturing
Qualcomm	QCS8250	CNN, GAN, RNN	TFL	smartphone, tablet, support 5G, video and image processing
Qualcomm	Snapdragon 888+	DNN	TFL	Smartphone, tablet, 5G, gaming, video upscaling, image & video processing,
RockChip	rk3399Pro	VGG16, ResNEt50, Inception4	TFL, Caffe, mxnet, ONNX, darknet	Smart Home, City, and Industry; face recognition, driving monitoring,
Rokid	Amlogic A311D	Inception V3, YoloV2, YOLOV3	TFL, Caffe2 Darknet	High-performance multimedia
Samsung	Exynos 2100	CNN	TFL	Smartphone, tablet, advanced image signal processing (ISP), 5G
Samsung	HBM-PIM	DNN see youtube to write on int	Pytorch, TFL	Supercomputer and AI application
Synopsis	EV7x	CNN, RNN, LSTM	OpenCV, OpenVX and OpenCL C, TFL, Caffe2	Robotics, autopilot car, vision, SLAM, and DSP algorithms
Syntiant	NDP100	DNN	TFL	Mobile phones, hearing equipment, smartwatches, IoT, remote controls
Syntiant	NDP101	CNN, RNN, GRU, LSTM	TFL	Mobile phones, smart homes, remote controls, smartwatches, IoT
Syntiant	NDP102	CNN, RNN, GRU, LSTM	TFL	Mobile phones, smart homes, remote controls, smartwatches, IoT
Syntiant	NDP120	CNN, RNN, GRU, LSTM	TFL	Mobile phones, smart home, wearables, PC,

 doi:10.20944/preprints202407.0025.v2

				IoT endpoints, media streamers, AR/VR
Syntiant	NDP200	FC, Conv, DSConv, RNN-GRU, LSTM	TFL	Mobile phones, smart homes, security cameras, video doorbells
Think Silicon	Nema PicoXS	DNN	----	Wearable and embedded devices
Tesla	FSD	CNN	Pytorch	Automotive
Verisilicon	VIP9000	All modern DNN	TF, Pytorch, TFL, DarkNet, ONNX	Can perform as intelligent eye and intelligent ear at the edge
Untether	TsunAImi	DNN, ResNet-50, Yolo, Unet, RNN, BERT, TCNs, LSTMs	TFL, Pytorch	NLP, Inference at the edge server or data center
UPMEM	UPMEM-PIM	DNN	-----	Sequence alignment: DNA or protein; Genome assembly: Metagenomic analysis

Memryx [297] released an inference processor, MX3. This processor computes deep learning models with 4, 8, or 16 bit weight and BF16 activation functions. MX3 consumes about 1 W of power and computes with 5 TFLOPS. This chip stores 10 million parameters on a die, and thus needs more chips for implementing larger networks.

MobileEye and STMicroelectronics released EyeQ 5 SoC for autonomous driving [193]. EyeQ 5 is 4 times faster than their earlier version, EyeQ 4. It can produce 2.4 TOPS/W and goes up to 24 TOPS with 10 W of power [194]. Recently, MobileEye has announced their next generation processor, EyeQ6, which is around 5x faster than EyeQ5 [195]. For INT8 precision, EyeQ5 performs 16 TOPS, and EyeQ6 shows 34 TOPS [196].

NXP introduced their edge processor i.MX 8M+ for the targeted applications in vision, multimedia, and industrial automations [86]. The system includes a powerful Cortex-A53 processor integrated with an NPU. The neural network performs 2.3 TOPS with 2W of power consumption. The neural computation supports INT16 precision [87]. NXP is scheduled to launch its next AI processor, iMX9, in 2023 with more features and efficiency [85].

NVIDIA released the Jetson Nano, which is able to run multiple applications in parallel, such as image classification, object detection, segmentation, and speech processing [197]. This developer kit is supported by the NVIDIA JetPack SDK and is able to run state-of-the-art AI models. The Jetson Nano consumes around 5-10 W of power and computes 472 GFLOPS in FP16 precision. The new version of Jetson Nano B01 can perform 1.88 TOPS 198].

NVIDIA released Jetson Orin, which includes specialized development hardware, AGX Orin. It is embedded with 32GB of memory, has a 12-core CPU, and can exhibit a computing performance of 275 TOPS while using INT8 precision [199]. The computing is powered by NVIDIA ampere architecture with 2048 cores, 64 tensor cores, and 2 NVDLA v2.0 accelerators for deep learning [199].

Qualcomm developed the QCS8250 SoC for intensive camera and edge applications [200]. This processor supports wifi and 5G for the IoTs. A quad hexagon vector extension V66Q with hexagon

 doi:10.20944/preprints202407.0025.v2

DSP is used for machine learning. An integrated NPU is used for advanced video analysis. The NPU supports INT8 precision and runs at 15 TOPS [201]. Qualcomm has released the Snapdragon 888+ 5G processor for use in smartphones. It takes the smartphone experience to a new level with AI-enhanced gaming, streaming, and photography [202]. It includes a 6th generation Qualcomm AI engine with the Qualcomm Hexagon780 CPU [203,204]. The throughput of the AI engine is 32 TOPS with 5 W of power consumption [203]. The Snapdragon 8 Gen2 mobile platform was presented at the HotChips 2023 conference and exhibited 60% better energy efficiency than the Snapdragon 8 in INT4 precision.

Samsung announced the Exynos 2100 AI edge processor for smartphones, smartwatches, and automobiles [205]. Exynos supports 5G network and performs on-device AI computations with triple NPUs. They fabricate using 5nm extreme UV technology. The Exynos 2100 consumes 20% lower power and delivers 10% higher performance than Exynos 990. Exynos 2100 can perform up to 26 TOPS, and it is 2 times more power efficient than the earlier version of Exynos [206]. A more powerful mobile processor, Exynos 2200, was released recently.

SiMa.ai [300] introduced the MLSoC for computer vision applications. MLSoc is implemented on TSMC 16nm technology. The accelerator can compute 50 TOPS while consuming 10 W of power. MLSoC uses INT8 precision in computation. The processor has 4 MB of on-chip memory for the deep learning operations. The processor is 1.4x more efficient than Orin, measured in frames/W.

Tsinghua and Polar Bear Tech released their QM930 accelerator consisting of seven chiplets. The chiplets are organized as one hub chiplet and six side chiplets, forming a Hub-Side processor. The processor is implemented in 12nm CMOS technology. The total area for the chiplets is 209 mm^2 for seven chiplets. However, the total substrate area of the processor is 1089 mm^2. The processor can compute with INT4, INT8, and INT16 precision, showing peak performances of 40, 20, and 10 TOPS, respectively. The system energy efficiency is 1.67 TOPS/W while computed in INT8. The power consumption can be varied from 4.5 to 12 W.

Verisilicon brought VIP 9000 for face and voice recognition. It adopts Vivante's latest VIP V8 NPU architecture for processing neural networks [207]. The computing engine supports INT8, INT16, FP16, and BF16. The performance can be scaled from 0.5 to 100 TOPS [208].

Synopsis developed the EV7x multi-core processor family for vision applications [209]. The processor integrates vector DSP, vector FPU, and a neural network accelerator. Each VPU supports a 32-bit scalar unit. The MAC can be configured for INT8, INT16, or INT32 precisions. The chip can achieve up to 2.7 TOPS in performance [210].

Tesla designed the FSD processor which was manufactured by Samsung for autonomous vehicle operations [211]. The SoC processor includes 2 NPUs and one GPU. The NPUs support INT8 precision, and each NPU can compute 36.86 TOPS. The peak performance of the FSD chip is 73.7 TOPS. The total TDP power consumption of each FSD chip is 36 W [211].

Several other companies have also developed edge processors for various applications but did not share hardware performance details on their websites or through publicly available publications. For instance, Ambarella [307] has developed various edge processors for automotive, security, consumer, and IoTs for industrial and robotics applications. Ambarella's processors are SoC types, mainly using ARM processors and GPUs for DNN computations.

ii. Neuromorphic Edge AI Processor

In 2022, the global market value of neuromorphic chips was 3.7 billion, and by 2028, the estimated market value is projected to be $ 27.85 Billion [212]. The neuromorphic processors described in this section utilize spike-based processing.

Synsense (formerly AICTx) has introduced a line of ultra-low power neuromorphic processors: DYNAP-CNN, XYLO, DYNAP-SE2, and DYNAP-SEL [15]. Of these, we were able to find performance information on only the DYNAP-CNN chip. This processor is fabricated on a 22 nm process technology and has a die area of 12 mm^2. Each chip can implement up to a million spiking neurons, and a collection of DYNAP-CNN chips can be utilized to implement a larger CNN architecture. The chip utilizes asynchronous processing circuits [213].

 doi:10.20944/preprints202407.0025.v2

BrainChip introduced the Akida line of spiking processors. The AKD1000 has 80 NPUs, 3 pJ/synaptic operation, and around 2 W of power consumption [147]. Each NPU consists of eight neural processing engines that run simultaneously and control convolution, pooling, and activation (ReLu) operations [148]. Convolution is normally carried out in INT8 precision, but it can be programmed for INT 1, 2, 3 or 4 precisions while sacrificing 1-3% accuracy. BrainChip has announced future releases of smaller and larger Akida processors under the AKD500, AKD1500, and AKD2000 labels [148]. A trained DNN network can be converted to SNN by using the CNN2SNN tool in the Meta-TF framework for loading a model into an Akida processor. This processor also has on-chip training capability, thus allowing the training of SNNs from scratch by using the Meta-TF framework [146].

GrAI Matters Lab (GML) developed and optimized a neuromorphic SoC processor named as VIP for computer vision application. VIP is a low power and low latency AI processor, with 5-10 W of power consumption, and the latency is 10x less than the NVIDIA nano [214]. The target applications are for audio/video processing on the end devices.

IBM developed the TrueNorth neuromorphic spiking system for real-time tracking, identification, and detection [10]. It consists of 4096 neurosynaptic cores and 1 million digital neurons. The typical power consumption is 65 mW, and the processor can execute 46 GSOPS/W, with 26 pJ per synaptic operation [10,215]. The total area of the chip is 430 mm^2, which is almost 14x bigger than that of Intel's Loihi 2.

Innatera announced a neuromorphic chip that is fabricated using TSMC's 28 nm process [216]. When tested with audio signals [217], each spike event consumed about 200 fJ, while the chip consumed only 100 uW for each inference event. The target application areas are mainly audio, healthcare, and radar voice recognition [217].

Intel released the Loihi [9], a spiking neural network chip in 2018, and an updated version, the Loihi 2 [9], in 2021. The Loihi 2 is fabricated using Intel's 7nm technology and has 2.3 billion transistors with a chip area of 31mm^2. This processor has 128 neuron cores and 6 low power x86 cores. It can evaluate up to 1 million neurons and 120 million synapses. The Loihi chips support online learning. Loihi processors support INT8 precision. Loihi 1 can deliver 30 GSOPS with 15 pJ per synaptic operation [218]. Both Loihi 1 and Loihi 2 consume similar amounts of power (110mW and 100mW, respectively [219]). However, the Loihi 2 outperforms the Loihi 1 by 10 times. The chips can be programmed through several frameworks, including, Nengo, NxSDK, and Lava [150]. The latter is a framework developed by Intel and is being pushed as the primary platform to program the Loihi 2.

IMEC develops a RISC-V processor based digital neuromorphic processor with 22nm process technology in 2022 [220]. They have implemented an optimized BF-16 processing pipeline inside the neural process engine. The computation can also support INT4 and INT8 precision. They have used 3-layer memory to reduce the chip area.

Koniku combines biological machines with silicon devices to design a micro electrode array system core [12]. They are developing the hardware and algorithm that mimics the smell sensory receptor that is found in some animal noses. However, the detailed device parameters are not publicly available. The device is mainly used in security, agriculture, and safe flight operation [221].

iii. PIM Processor

PIM processors are becoming an alternative for AI application due to the low latency, high energy efficiency, and reduced memory requirements. PIMs are the analog, and in-place computing architecture thus, it reduces the burden of additional storage modules. However, there are some digital presents the schematic representation of a common PIM computing architecture. It consists of the crossbar array (NxM) of the popular storage devices. The crossbar array performs as the weight storage and analog multiplier. The storage devices could be SRAM, RRAM, PCM, STT-MRAM or a flash memory cell. The computing array is equipped with the peripheral circuits, a data converter (ADC or DAC), sensing circuits, and a write circuit for the crossbar. Some of the PIM processors are discussed in this section.

 doi:10.20944/preprints202407.0025.v2

Imec and GlobalFoundries have developed DIANA, a processor that includes both digital and analog cores for DNN processing. The digital core is employed for widely parallel computation, whereas the analog in-memory computing (AiMC) core enables much higher energy efficiency and throughput. The core uses a 6T-SRAM array with a size of 1152x512. Imec developed the architecture, while the chip is fabricated using GlobalFoundries' 22FDX solution [222]. It is targeted for a wide range of edge applications, from smart speakers to self-driving vehicles. The analog component (AiMC) computes at 29.5 TOPS with and the digital core computes at 0.14 TOPS. The digital and analog components have efficiencies of 4.1 TOPS/W and 410 TOPS/W, respectively in isolation. The overall system energy efficiency of DIANA is 14.4 TOPS/W for Cifar-10 [223].

Gyrfalcon has developed multiple PIM processors, including the Lightspeeur 5801, 2801, 2802, and 2803 [24]. The architecture uses digital AI processing in-memory units that compute a series of matrices for CNN. The Lightspeeur 5801 has a performance of 2.8 TOPS at 224 mW and can be scaled up to 12.6 TOPS/W. The Lightspeeur 2803S is their latest PIM processor for the advanced edge, desktop, and data center deployments [19]. Each Lightspeeur 2803S chip performs 16.8 TOPS while consuming 0.7W of power, giving an efficiency of 24 TOPS/W. Lightspeeur 2801 can compute 5.6 TOPS with an energy efficiency of 9.3 TOPS/W. Gyrfalcon introduced its latest processor, Lightspeeur 2802, using TSMC's magnetoresistive random access memory technology. Lightspeeur 2802 exhibits an energy efficiency of 9.9 TOPS/W. Janux GS31 is the edge inference server which is built with 128 Lightspeeur 2803S chips [225]. It can perform 2150 TOPS and consumes 650 W.

Mythic has announced its new analog matrix processor, M1076 [18]. The latest version of Mythic's PIM processor reduced its size by combining 76 analog computing tiles, while the original one (M1108) uses 108 tiles. The smaller size offers more compatibility to implant on edge devices. The processor supports 79.69M on-chip weights in the array of flash memory, and 19,456 ADCs for parallel processing. There is no external DRAM storage required. The DNN models are quantized from FP32 to INT8 and retrained in Mythic's analog compute engine. A single M1076 chip can deliver up to 25 TOPS while consuming 3 W of power [90]. The system can be scaled for high performance up to 400 TOPS by combining 16 of M1076 chips which requires 75 W [88,89].

Samsung has announced its HBM-PIM machine learning-enabled memory system with PIM architecture [16]. This is the first successful integration of the PIM architecture of high bandwidth memory. This technology incorporates the AI processing function into Samsung HBM2 Aquabolt to speed up high-speed data processing in supercomputers. The system delivered 2.5x performance with 60% lower energy consumption than the earlier HBM1 [16]. Samsung LPDDR5-PIM memory technology for mobile device technology is targeted to bring the AI capability in the mobile device without connecting to the data center [226]. The HBM-PIM architecture is different from the traditional analog PIM architecture as outlined in Figure 2. It does not require data-conversion and sensing circuits as the actual computation is taking place in the near-computing module in the digital domain. Instead, it places a GPU surrounded by HBM stacks to realize the parallel processing and minimize data movement [227]. Therefore, this is similar to a dataflow processor

Syntiant has developed a line of flash-memory array based edge inference processors, such as NDP10x, NDP120, NDP200 [228]. Syntiant's PIM architecture is very energy efficient, and it combines with an edge optimized training pipeline. A Cortex-M0 is embedded in the system that runs the NDP firmware. The NDP10x processors can hold 560k weights of INT4 precision and perform MAC operation with an INT8 activation. The training pipeline can build neural networks for various applications according to the specifications with optimized latency, memory size, and power consumption [228]. Syntiant released five different versions of application processors. NDP 100 is their first AI processor, updated in 2020 with a tiny little dimension of 2.52 mm^2 and ultra-low power consumption, less than 140 μW [229]. Syntiant continues to provide more PIM processors named NDP 101, 102, 120, and NDP 200 [230–232]. The application domains are mainly smartphone, wearable, and hearable pieces of equipment, remote controls, IoT endpoints. The neural computations are supported by INT 1, 2, 4, and 8 precision. The energy efficiency of the NDP 10x series is 2 TOPS/W [233], which includes NDP100, NDP 101, and NDP 102. NDP 120 [235] and NDP 200 exhibit 1.9 GOPS/W and 6.4 GOPS/W [232], respectively.

 doi:10.20944/preprints202407.0025.v2

Untether has developed its PIM AI accelerator card TsunAImi [236] for inference at the data center or in the server. The heart of the TsunAImi is four runAI200 chips which are fabricated by TSMC in standard SRAM arrays. Each runAI200 chip features 511 cores and 192MB of SRAM memory. runAI200 computes in INT8 precision and performs 502 TOPS at 8 TOPS/W, which is 3x more than NVIDIA's Ampere A100 GPU. The resulting performance of TsunAImi system is 2008 TOPS with 400 W [237].

UPMEMP PIM innovatively placed thousands of DPU units within the DRAM memory chips [238]. The DPUs are controlled by high-level applications running on the main CPU. Each DIMM consists of 16 PIM-enabled chips. Each PIM has 8 DPUs, making 128 DPUs of total DPUs for each UPMEM [239].

However, the system is massively parallel, and up to 2560 DPUs units can be assembled as a unit server with 256 GB PIM DRAM. The computing power is 15x of x86 server with the main CPU. The throughput is benchmarked for INT32 bit addition is 58.3 MOPS/DPU [240]. This system is suitable for DNA sequencing, Genome comparison; Phylogenetics; Metagenomic analysis, and more [241].

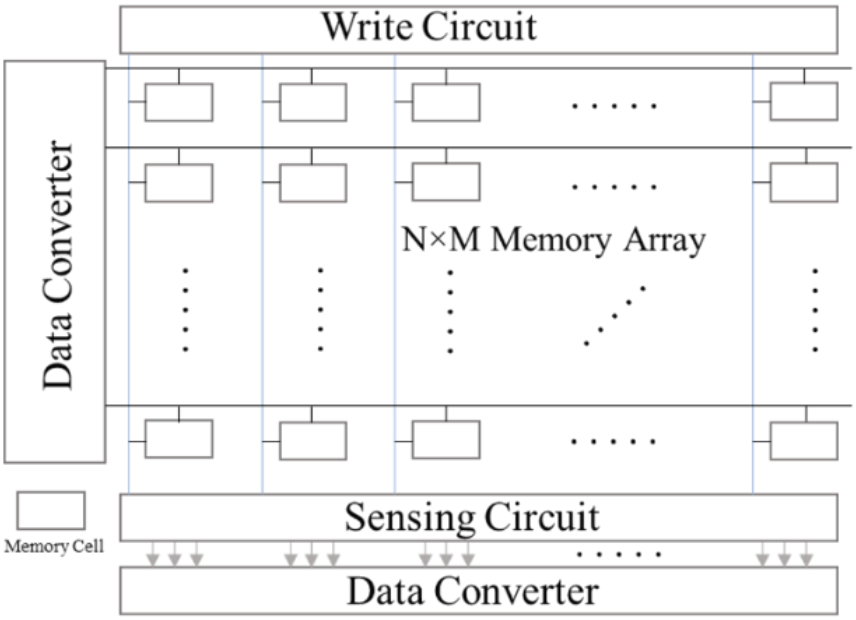

Figure 2. Schematic representation of a Processing in Memory macro system.

iv. Processor in the Industrial Research

The PIM computing paradigm is still in its rudimentary stage, however a very promising system for efficient MAC operation and low power edge application. A good number of industries and industry-academic research collaborations are escalating the development of PIM technologies and architectures. In this section, the PIM processors in the industrial and industry-university collaboration have been briefly discussed. The recent developments of the PIM research are tabulated in Table 4.

 doi:10.20944/preprints202407.0025.v2

Table 4. Edge Processors in Industrial Research with Technology, Process Technology, and Numerical Precision.

Research Group	Name	Power (W)	Process (nm)	Area (mm^2)	Precision INT/FP*	Performance (TOPS)	E. Eff. (TOPS/W)	Architecture	Reference
TSMC+ NTHU		2.13E-03	22	6	2,4,8	4.18E-01	195.7	PIM	[259]
TSMC		0.037	22	0.202	4,8,12,16	3.3	89	PIM	[257]
TSMC		0.00142	7	0.0032	4	0.372	351	PIM	[258]
Samsung+GIT	FORMS	66.36	32	89.15	8	0.0277		PIM	[262]
IBM + U Patra	HERMES	9.61E-02	14	0.6351	8	2.1	21.9	PIM	[260]
Samsung+ASU	PIMCA	0.124		20.9	1,2	4.9	588	PIM	[261]
Intel+Cornell U	CAPE		7	9	4			PIM	[263]
SK Hynix	AiM			6.08		1		PIM	[264]
TSMC	DCIM	0.0116	5	0.0133	4,8	2.95	254	PIM	[265]
Samsung		0.3181	4	4.74	4,8,16, FP16	39.3	11.59	Dataflow	[266]
Alibaba + FU		0.0212	28	8.7	3	0.97	32.9	Dataflow	[267]
Alibaba + FU		0.072	65	8.7	3	1	8.6	Dataflow	[267]
Alibaba		0.978	55	602.22	8			Dataflow	[268]
TSMC+ NTHU		0.00227	22	18	2,4,8	0.91	960.2	PIM	[269]
TSMC + NTHU		0.00543	40	18	2,4,8	3.9	718	PIM	[270]
TSMC+GIT		0.000350	40	0.027		0.0092	26.56	PIM	[271]
TSMC+GIT		0.131	40	25	1-8,1-8,32	7.989	60.64	PIM	[272]
Intel+UC		0.0090	28	0.033	1,1	20	2219	PIM	[273]
Intel+UC		0.0194	28	0.049	1-4,1	4.8	248	PIM	[274]
TSMC+ NTHU	nvCIM	0.00398	22	6	2,4	5.12	1286.4	PIM	[275]
Pi2star +NTHU		0.00841.	65	12	1-8	3.16	75.9	PIM	[276]
Pi2star +NTHU		0.00652	65	9	4,8	2	35.8	PIM	[277]
Tsing+NTHU		0.273	28	6.82	12	4.07	27.5	Dataflow	[278]
Samsung		0.381	4	4.74	4,8,FP16	19.7	11.59	Dataflow	[279]
Renesas Electronics		4.4	12			60.4	13.8	Dataflow	[280]
IBM		6.20	7	19.6	2,4,FP(8,16,32)	102.4	16.5	Dataflow	[281]
Intel + IMTU	QNAP	0.132	28	3.24	8	2.3	17.5	Dataflow	[282]
Samsung		0.794	5	5.46	8,16	29.4	13.6	Dataflow	[283]
Sony		0.379	22	61.91	8,16,32	1.21	4.97	Dataflow	[284]
Mediatek		1.05	7	3.04		3.6	13.32	Dataflow	[285]
Pi2star		0.099	65	12	8	1.32	13.3	Dataflow	[286]
Mediatek		0.0012	12			0.102	86.24	PIM	[287]
TSMC+NTHU		0.10	22	8.6	8,8,8	6.96	68.9	PIM	[288]
TSMC+NTHU		0.099	22	9.32	8,8,8	24.8	251	PIM	[289]
ARM+Harvard		0.04	12		FP4	0.734	18.1	Dataflow	[290]
ARM+Harvard		0.045	12		FP8	0.367	8.24	Dataflow	[291]
TSMC + NTHU		0.0037	22	18	8,8,22	0.59	160.1	Dataflow	[292]

 doi:10.20944/preprints202407.0025.v2

STMircroelectronics		0.738	18	4.24	1,1	229	310	Dataflow	[293]
STMircroelectronics		0.740	18	4.19	4,4	57	77	Dataflow	[294]
MediaTek		0.711	12	1.37	12	16.5	23.2	PIM	[309]
TSMC+ NTHU			16		8		98.5	PIM	[308]
Renesas Electronics		5.06	14		8	130.55	23.9	Dataflow	[310]

* Integer Precision is indicated by only precision number(s). Floating point precision is mentioned as FP in the precision column.

Alibaba has developed SRAM and DRAM-based digital CIM and PNM systems for low precision edge applications [267,268]. The CIM architecture uses multiple chiplet modules (MCM) to solve the complex problem instead of a single SoC. The CIM architecture in [267] proposes Computing-on-Memory Boundary (COMB), which is a compromise between in-memory and near-memory computation. This technique exhibits high macro computing energy efficiency and low system power overhead. This CIM architecture demonstrated scaleable MCM systems using a COMB NN processor. The layerwise pipeline mapping schemes are utilized to deploy different sizes of NN for the required operation. The chip operation is demonstrated with keyword spotting, CIFAR-10 image classification, and object detection with tiny-YOLO NN using one, two, and four chiplets.

IBM and the University of Patra together presented their PCM-based CIM processor, HERMES [260]. This CIM is a 256×256 in-memory compute core fabricated in 14nm CMOS process technology for edge inference. HERMES is demonstrated for image classification operation on MNIST and CIFAR-10 datasets.

Samsung technology has been working on various CIM architectures for AI applications in edge to the datacenter. The company has released HBM-PIM recently [226,227]. HBM-PIM is for high-speed memory access, which is fabricated with DRAM in a 20nm process. Samsung and Arizona State University (ASU) presented a PIMCA chip for AI inference [261]. PIMCA chip for AI inference [261]. PIMCA consumes a very low amount of power (124 mW). PIMCA is highly energy efficient, 588 TOPS/W as shown in Table 2. TSMC has designed and fabricated analog [257,258] and digital [265] CIM systems for inference.

Besides TSMC's own research, the company has multiple CIM research projects on various emerging memory devices such as ReRAM [259], STT-MRAM [269], PCM [270], RRAM [271] and RRAM-SRAM [272] in collaboration with various research groups in the academia. The performance of these macro inference chips has been demonstrated in various high-tier conferences or scientific forums very recently. The best performance is demonstrated in ISSCC 2022 with PCM devices, and it exhibits 5.12 TOPS in 2-bit precision [275], which is 1286.4 TOPS/W. This CIM processor supports INT2 and 4 bit computing precision. The digital CIM system is fabricated with FinFET in 5nm process technology, and it performs 2.95 TOPS and 254 TOPS/W [265].

Besides the AI accelerators introduced above, there are a handful of companies that are working on edge processors. The companies working on neuromorphic processors are MemComputing [213,295], GrAI [214], and iniLabs [296]. Memryx is a recently formed startup which is building high performance and energy efficient AI processors for a wide range of applications, such as transportation, IoT, and industry [297]. It can compute Bfloat16 activation with 4/8/16-bit weight and performs about 5TFLOPS.

7. Performance Analysis of the Edge Processors

This section discusses the performance analysis of the edge processors described earlier. The discussion is focused on different architectures for edge processors. At first, the overall performance is discussed based on the computing performance, power consumption, chip area, and computing precision. Then only PIM processors are discussed. At the end of this section, we have focused on the devices still under research and development or waiting for commercially available.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0025.v2

i. Overall Analysis of the AI Edge Processor

We compare all the edge AI processors listed in the previous section using the following key metrics:

1. Performance: tera-operations per second (TOPS).
2. Energy efficiency: TOPS/W.
3. Power: Watt (W).
4. Area: square millimeter (mm^2).

Performance: Figure 3 plots the performance vs. power consumption, with different labels for dataflow, neuromorphic, and PIM processors. The processors within a power consumption range of 1W to 60W have a performance of 1 to 275 TOPS. These are geared towards comparatively high-power applications such as surveillance systems, autonomous vehicles, industries, smart cities, and UAVs. The highest throughput processors in this list are the EyeQ6 from MobileEye, the Journey 5 from Horizon, and the Jetson Orin from Nvidia. The Jetson Orin is about 2.15 times faster than both the EyeQ6 and Journey 5. From the company datasheet [], the Jetson Orin has 275 TOPS at INT8 precision for 60W of power. The Orin consumes about 1.5 and 2 times more power than the EyeQ6 and Journey 5, respectively. The processors with a power consumption of less than 1W have a performance from 0.2 GOPS to 17 TOPS.

The IBM NorthPole has 200 TOPS for INT8 precision at 60 W (based on a discussion with IBM). However, the NorthPole can have higher TOPS of 400 and 800 at 4 and 2 bit precision respectively. According to a recent NorthPole article, the maximum power consumption of the NorthPole processor is 74 W [304].

These are targeted for extreme edge and IoT applications. The least power is needed for PIM processors of the NDP series by Syntiant, which are flash-memory based PIM processors [20].

Among neuromorphic processors, Loihi 2 outperforms other neuromorphic processors, except for the Akida AKD1000. The AKD1000 however consumes 20x more power than the Loihi 2 (see Table 2). Although the neuromorphic processors seem less impressive in terms of TOPS vs. W, it is important to note that they generally need far fewer synaptic operations to perform a task, if the task is performed with an algorithm that is natively spiking (i.e., not a deep network implemented with spiking neurons) [298].

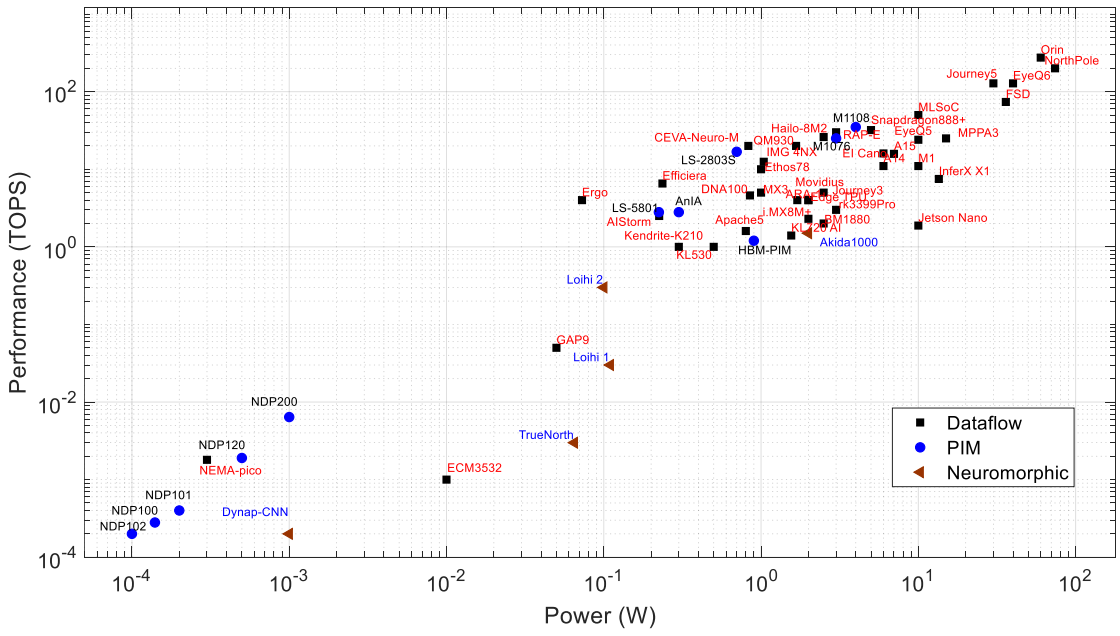

Figure 3. Power consumption and performance of the AI edge processors.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0025.v2

The neuromorphic processors consume significantly less energy than other processors for inference tasks [142]. For example, the Loihi processor consumes 5.3x less energy than the Intel Movidious and 20.5x less energy than the Nvidia Jetson Nano [142]. Figure 3 shows that higher performance PIM processors (such as the M1076, M1108, LS-2803S, and AnIA), exhibit similar computing speeds as dataflow or neuromorphic processors at the same range of power consumption (0.5 to 1.5 W).

Precision: Data precision is an important consideration when comparing processor performances. Figure 5 presents the precision of the processors from Figure 3. Figure 5 shows the distribution of precision and total number of processors for each architecture category. A processor may support more than one type of computing precision. Figures 3 and 4 are based on the highest precision supported for each processor.

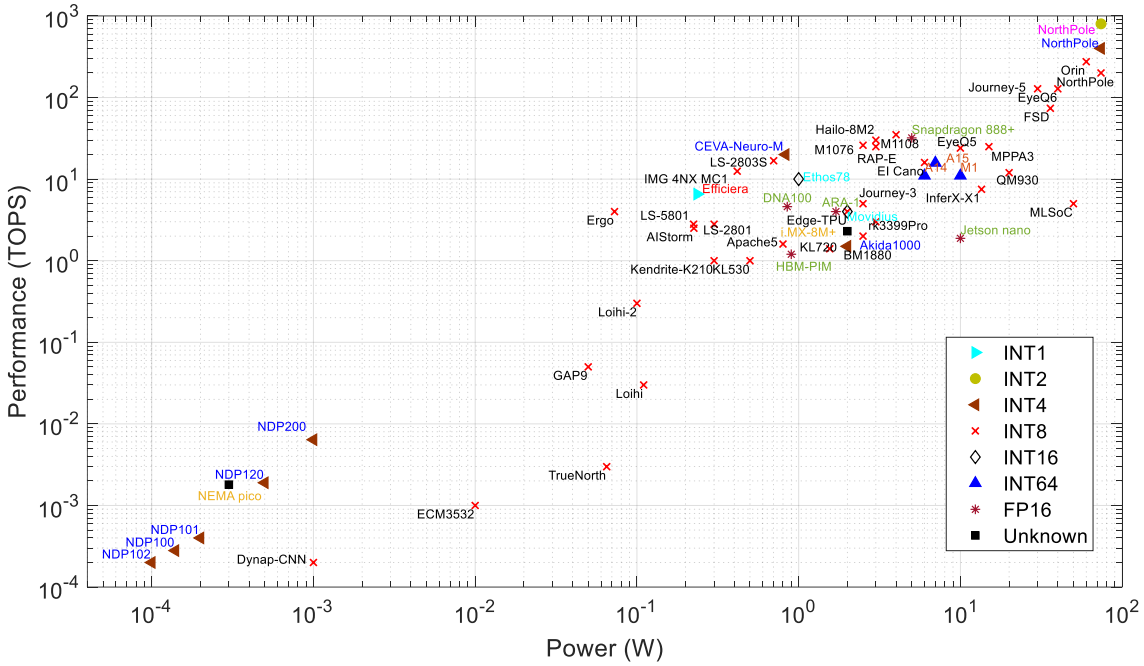

Figure 4. Power vs. performance of edge processors with computing precision.

Among dataflow processors, INT8 is the most widely supported precision for DNN computations. NVIDIA's Orin achieves 275 TOPS with INT8 precision, the maximum computing speed for INT8 precision in Figure 5. However, some processors utilize INT1 (Efficiara), INT64 (A15, A14, and M1), FP16 (ARA-1, DNA100, Jetson Nano, Snapdragon 888+), and INT16 (Ethos78, and Movidius). Neuromorphic and PIM processors mainly support INT1 to INT8 data precisions. Lower computing precisions generally reduce the inference accuracy. According to [185], VGG-9 and ResNet-18 have accuracy losses of 3.89% and 6.02%, respectively, for inference while computed with INT1 precision. A more in-depth discussion of the relationship between quantization and accuracy is presented in Section 3(i). A higher precision provides better accuracy but incurs more computing costs. Figure 5 shows that the most common precision in the processors examined is INT8. This provides a good balance between accuracy and computational costs.

 doi:10.20944/preprints202407.0025.v2

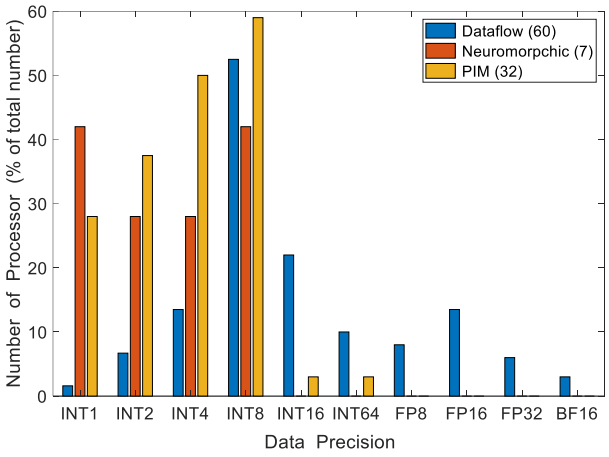

Figure 5. Number of edge processors supporting various data precision. Total number of processors is indicated in the legend.

As shown in Figure 4, and 5, almost all the neuromorphic processors use INT8 for synaptic computations. The exception to this is the AKD1000, which uses INT4 and shows the best performance among neuromorphic processors in terms of operations per second (1.5 TOPS). It however consumes around 18x more power than Loihi processors. At INT8 precision, the Loihi 1 performs 30 GSOPS using 110 mW [218,219], whereas Loihi 2 surpasses this throughput by 10x, with a similar power consumption [9].

As shown in Figure 4,5 PIM processors primarily support precisions of INT1 to INT8. Figure 5 shows the performance of PIM processors in INT4 and INT8 precisions due to the unavailability of data for all supported precisions. Mythic processors (M1108 and M1076) manifest the best performance among PIM processors. Mythic and Syntiant have developed their PIM processors with flash-memory devices. However, Mythic processors require significantly higher power to compute DNNs in INT8 precision with its 76 computing tiles. Syntiant processors use INT4 precision and compute with about 13,000x lower throughput than Mythic M1076 while consuming about 6000x less power. The Syntiant processors are limited to smaller networks with up to 64 classes in NDP10x. On the other hand, Mythic processors can handle 10x more weights with greater precision [233]. The Samsung DRAM architecture-based PIM processor uses computing modules near the memory banks and supports INT64 precision [16].

Energy Efficiency: Figure 6 presents the performance vs. energy efficiency of dataflow, PIM and neuromorphic processors. The efficiency determines the computing throughput of a processor per watt. The energy efficiency of all PIM processors is located within 1 to 16 TOPS/W, whereas most of the dataflow processors are located in 0.1 to 55 TOPS/W. The PIM architecture reduces latency by executing the computation inside the memory modules, which increases computing performance and reduces power consumption. Loihi 2 manifests best energy efficiency among all neuromorphic processors. Energy Efficiency vs. power consumption, as shown in Figure 7, gives us a better understanding about the processors. Loihi 2 shows better energy efficiency than many high performances edge AI processors while it consumes very low power. Ergo is the most energy efficient processor among all dataflow processors, which shows 55 TOPS/W. Ergo is the most energy efficient processor among all dataflow processors, which processors, which shows 55 TOPS/W.

Chip Area: The area is an important factor for choosing a processor for AI applications on edge devices. Modern processor technologies are pushing the boundaries to fabricate very high density and superior performance at the same time. The smaller die area and lower power consumption is very important for battery powered edge devices. The chip area is related to the cost of the silicon fabrication and also defines the application area. A smaller chip with high performance is desirable for edge applications.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0025.v2

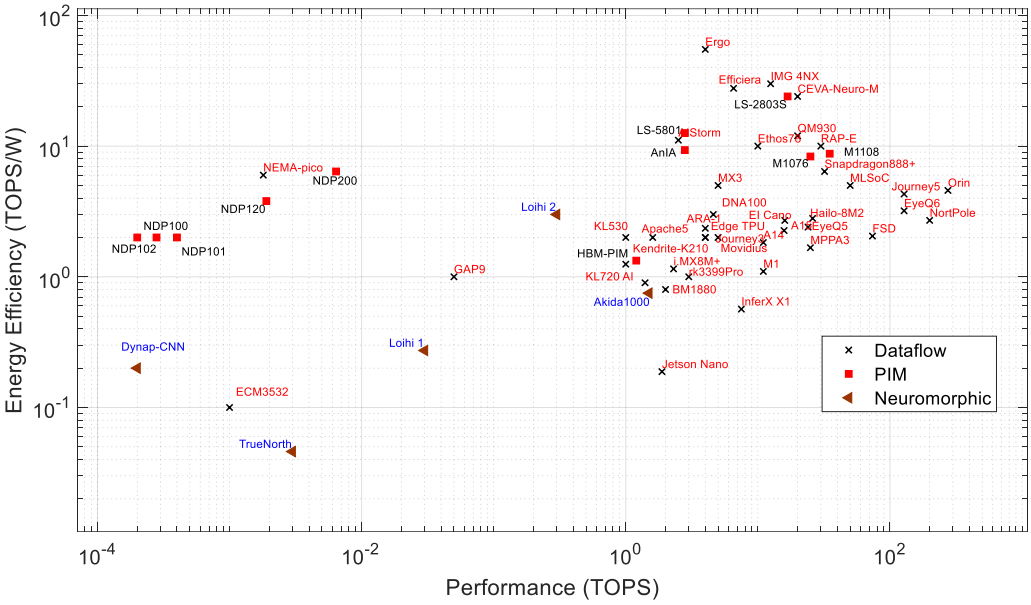

Figure 6. Performance and energy efficiency of the edge processors.

Figures 7 and 8 present the power consumption and performance, respectively vs. the chip area. It can be observed that in general, both the power consumption and performance increase with chip area. Based on the available chip sizes, the NothPole has the largest chip size of 800 mm² and performs 200 TOPS in INT8. The lowest area chips have a dataflow architecture. Figure 9 shows the energy efficiency vs. area as the combined relationship of Figures 8, and 9. In this Figure, the PIM processors form a cluster. The overall energy efficiency of this PIM cluster is higher than dataflow and neuromorphic processors of similar chip area. Some of the dataflow processors (such as Nema Pico, Efficiera, and IMG 4NX) exhibit higher energy efficiency and better performance vs area than other processors.

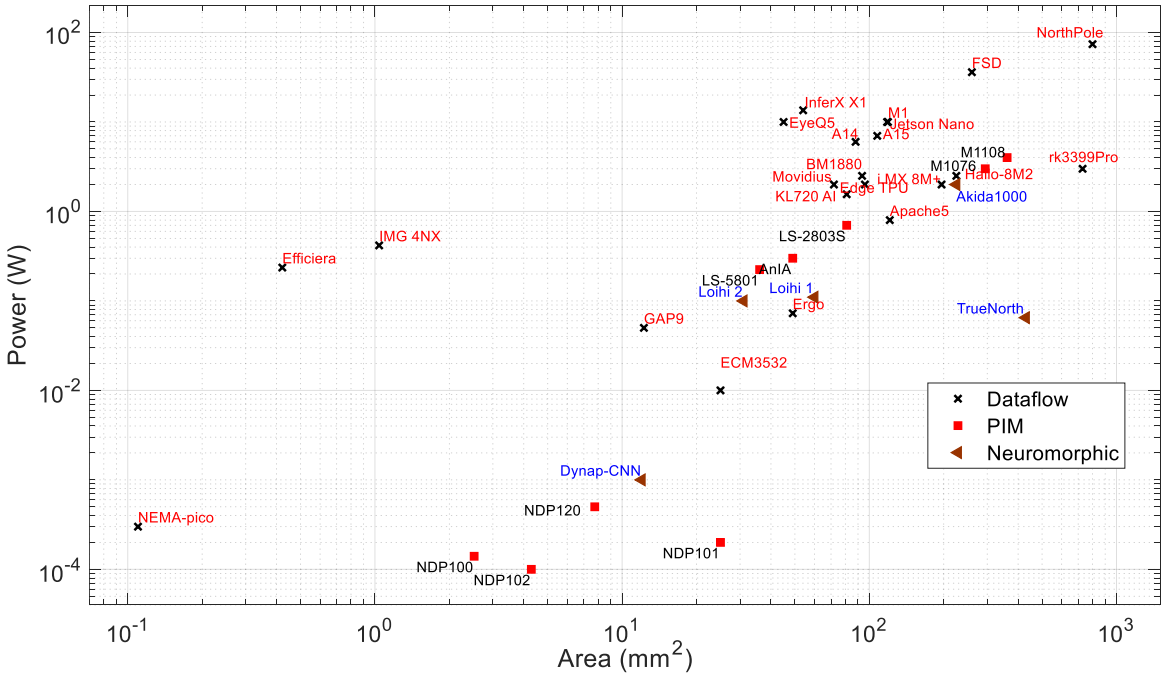

Figure 7. Power consumption vs. area of the edge pro.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0025.v2

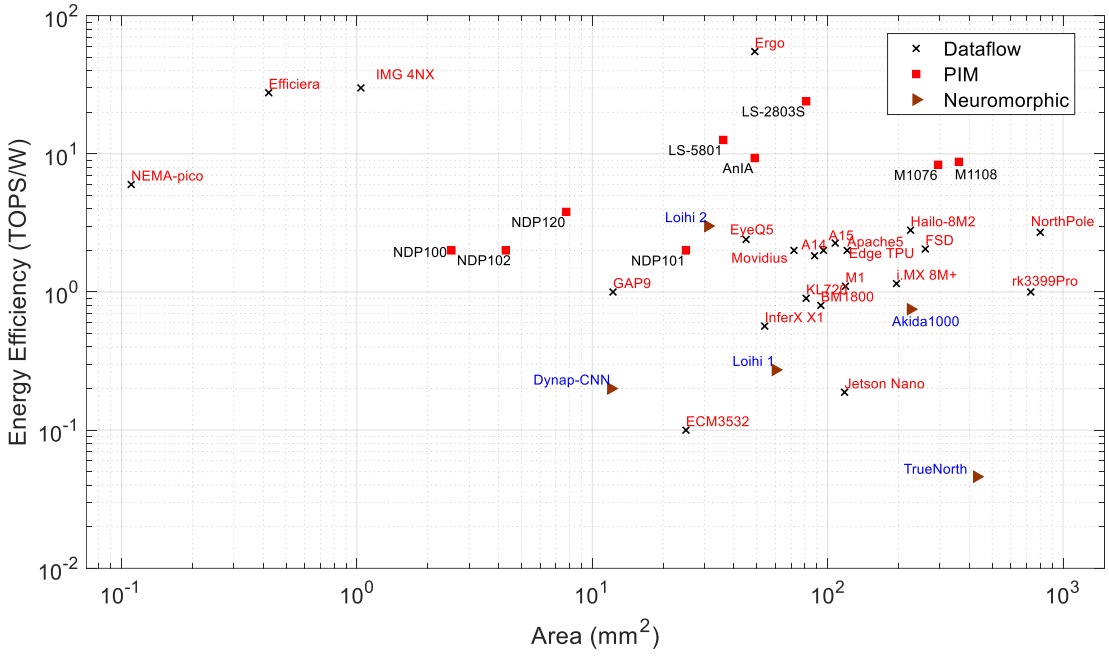

Figure 8. Area vs. performance of the edge processors.

ii. AI Edge Processor with PIM Architecture

While Figures 3–11 describe processors of all types, Figure 12 shows the relation between only PIM processors that are either announced as products or are still in the industrial research. The research processors are presented in the conferences, such as ISSCC and VLSI. The PIM processors at the lower right corner of Figure 11 are candidates for data center and intensive computing applications [236–241]. The PIM processors with higher energy efficiency are suitable for edge and IoT applications because of their smaller size, lower power consumption, and higher energy efficiency. From Figure 12 we can see that most of the PIM processors under industrial research show higher energy efficiency than announced processors. This indicates that future PIM processors are likely to have much better performances and efficiencies.

The PIM processors compute the MAC operation inside the memory array thus reducing the data transfer latency. Generally, PIM processors compute in lower integer/fixed-point precision. A PIM processor generally supports INT 1-16 precision. However, according to our study, we found around 59% of the PIM processors support INT8 precision for MAC operation, as shown in Figure 5. The low precision computation is faster and requires lower power consumption compared to dataflow processors. The PIM edge processors consume 0.0001 to 4 W for deep learning inference applications, as presented in Table 2 and Figure 3. However, the dataflow processors suffer from high memory requirements and latency issues and consume higher power than most of the PIM processors to get the same performance as we see in Figure 3-5.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0025.v2

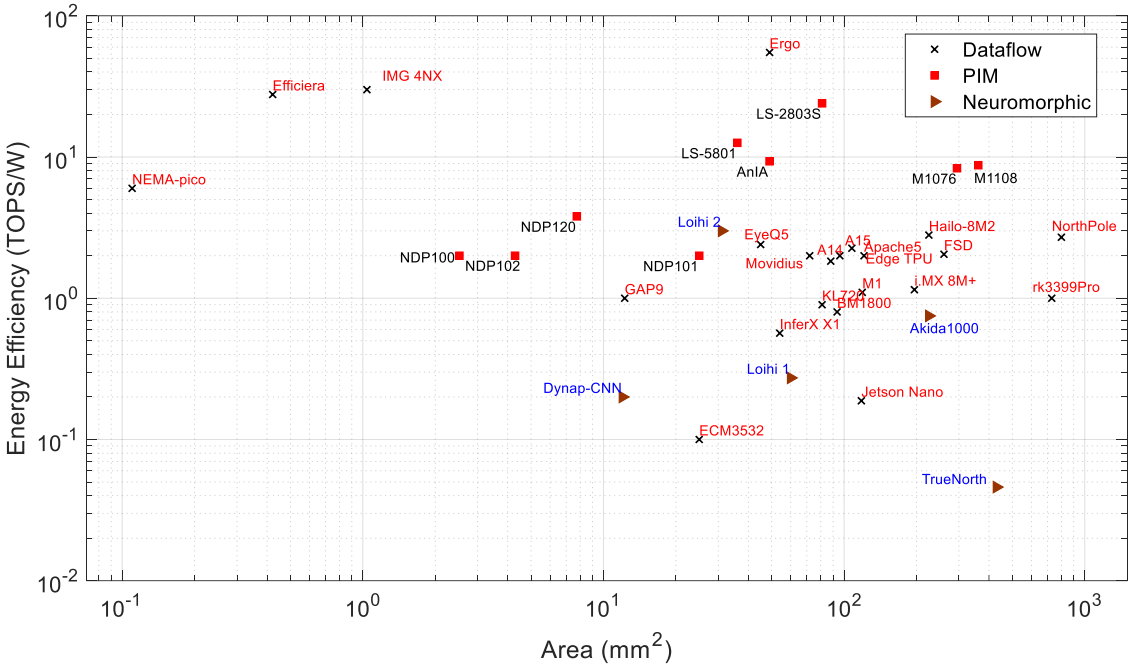

Figure 9. Area vs. energy efficiency of edge processors.

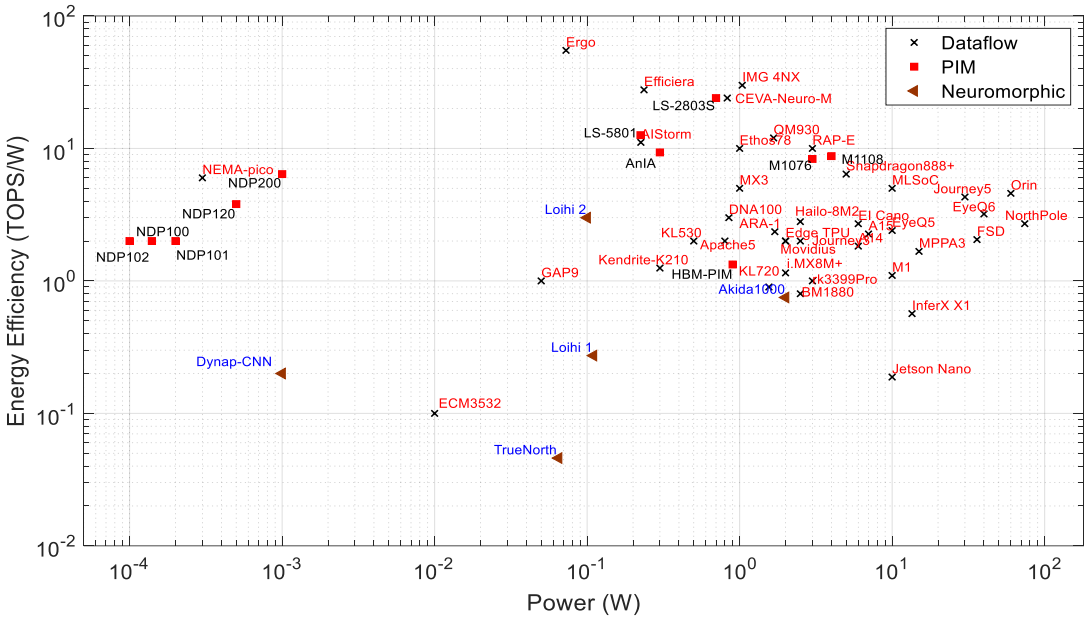

Figure 10. Power vs. energy efficiency of the edge processors.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 July 2024 doi:10.20944/preprints202407.0025.v2

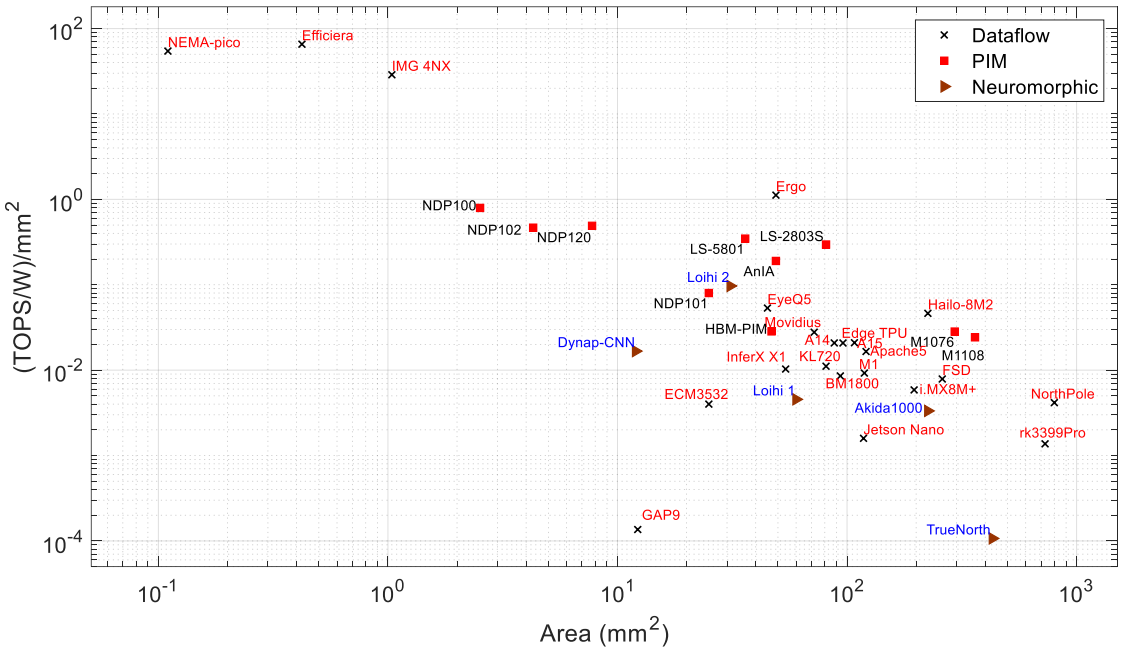

Figure 11. Area vs. energy efficiency per unit area of edge processors.

From Figure 3-4, Syntient's NDP200 consumes below 1mW power and shows the highest performance for extreme edge applications. Mythic M1108 consumes 4W and exhibits the highest performance (35 TOPS) than all dataflow and neuromorphic processors that consumes below 10 W of power. For the same chip area M1108 consumes 9x less power than Tesla's dataflow processor FSD, while FSD computes 2x faster than M1108 as presented in Figure 8-9.

The processors below 100 mm², Gyrfalcon's LS2803 shows the highest performance except EyeQ5. However, EyeQ5 consumes about 14x higher power and performs 1.4x better than LS2803. The benefit of deploying PIM processors for edge applications is its high performance with less power consumption, and the PIM processors reduce the computing latency significantly as the MAC operations perform inside the memory array.

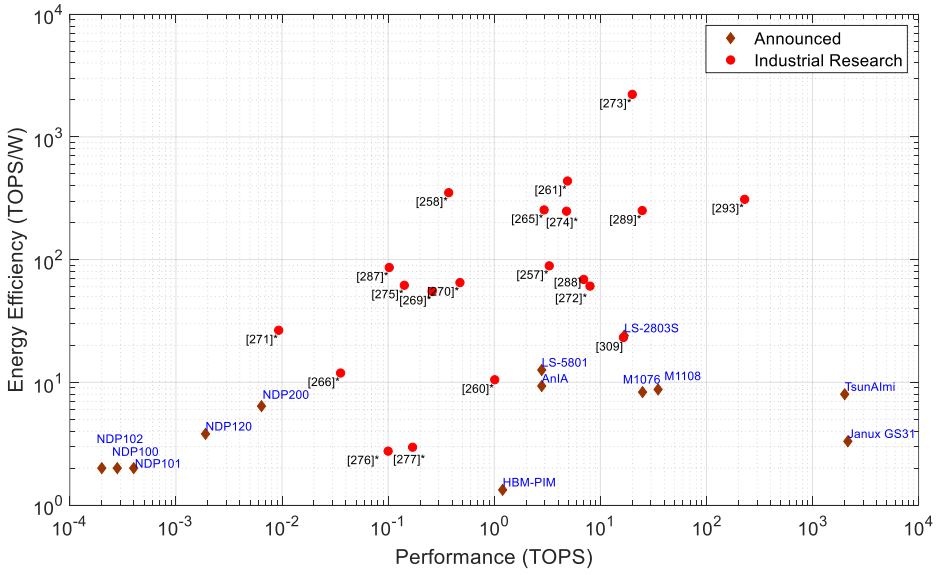

Figure 12. Performance vs. energy efficiency of PIM/CIM processors, Processors with Asterix (*) indicate the processors are still in industrial research, and other processors are released or announced by the manufacturer.

 doi:10.20944/preprints202407.0025.v2

iii. Edge Processor in the Industrial Research

Several companies, along with their collaborators, are developing edge computing architectures and infrastructures with state-of-the-art performance. Figure 13 shows the power consumption vs. energy efficiency of the industrial research processors which were presented at high tier conferences (such as ISSCC, VLSI). The chart includes both PIM [259–265,269–277,287–289] and dataflow [266-268,278-286.290-294] processors.

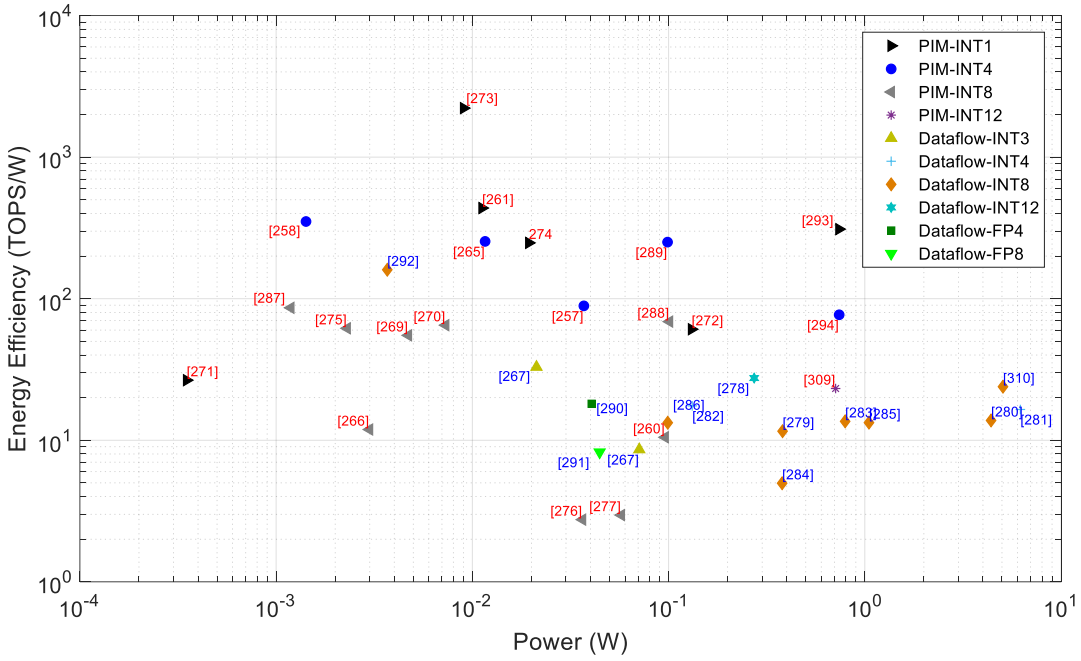

Figure 13. PIM (red) and dataflow (blue label) processors in industrial research.

Renesas Electronics presented a near-memory system in ISSCC 2024 developed in a 14nm process that achieved 130.55 TOPS with 23.9TOPS/W [310]. TSMC and National Tsing Hua University presented one near memory system in a 22nm CMOS process in ISSCC 2023 that computes 0.59 TOPS and 160 TOPS/W in 8-8-26 bit (input-weight-output) precision [292]. This system showed the highest energy efficiency amongst the near-memory and dataflow processors. The energy efficiency is achieved by a 90% input sparsity, while a 50% input sparsity gives an energy efficiency of 46.4 TOPS/W [292]. Alibaba and Fudan University [267] presented a processing near memory system in ISSCC 2022 with 0.97 TOPS and 32.9 TOPS/W energy efficiency while computing with INT3 precision. This accelerator is a SRAM based near-memory computing architecture. Tsing Microelectronics and Tsinghua University [278] demonstrated a dataflow processor in ISSCC 2022 for NLP and computer vision applications, which shows an energy efficiency of 27.5 TOPS/W in INT12 precision. Renesas Electronics [280] exhibited 13.8 TOPS/W in INT8 computing precision. Many other companies, such as IBM [281], Sony [284], Mediatek [285], and Samsung [279,283] have also demonstrated their research work on the dataflow edge processors with energy efficiencies around 11 to 18 TOPS/W

PIM processors generally manifest better energy efficiencies than dataflow processors. TSMC and National Tsing Hua University presented a PIM system in a 16nm CMOS process in ISSCC 2024 that achieved 98.5 TOPS/W in 8-8-23 precision (input-weight-output) [308]. Mediatek and TSMC presented a digital PIM system in ISCC 2024 developed in a 3nm process that achieved 23.2 TOPS/W with 16.5 TOPS performance [309]. Intel and Columbia University demonstrated a PIM processor [273] in ISSCC 2022 that shows the performance and energy efficiency of 2219 TOPS/W and 20 TOPS respectively, which is around 33x more efficient than the processor mentioned in [269].

However, the former processor uses extremely lower precision (INT1). TSMC and Tsinghua University [288] presented a PIM accelerator in ISSCC 2023 with 6.96 TOPS and 68.9 TOPS/W, which

 doi:10.20944/preprints202407.0025.v2

is about 12x faster than the near memory computing system presented in [292], while computing in INT8 precision. STMicroelectronics presented a PIM accelerator that computes 57 TOPS and 77 TOPS/W in INT4 precision [294] and performs about 25x better than near memory computing presented in ISSCC 2021 [282]. TSMC and Tsinghua University [270] presented a PCM based processor in ISSCC 2022, which shows 65 TOPS/W in INT8 precision, and is around 5x better than [292]. Samsung and Arizona State University [261] demonstrated PIMCA in VLSI' 2021 and showed an energy efficiency of 437 TOPS/W computed in INT1 precision. Other companies such as TSMC and collaborators [259,269–272,275,288,292], Samsung and collaborators [261], Intel and collaborators [263,273,274,282], and HK Hynix [264] have demonstrated their PIM processors in recent ISSCC and VLSI conferences.

8. Summary

This article reviewed different aspects and paradigms of recent AI edge processors released or announced recently by various tech companies. About 100 edge processors were examined. This work however did not cover the DNN algorithms, HPC computing processors, or cloud computing. We categorized state-of-the-art edge processors and analyzed their performance, area, and energy efficiency to support the research community in edge computing. Multiple processing architectures including dataflow, neuromorphic, and PIM were examined. The performance and power consumption were analyzed for narrowing down the edge AI processors for specific applications. The deep neural networks and software frameworks supported were discussed and presented in tables.

Several of the edge processors offer on-chip retraining in real-time. This enables retraining of networks without having to send sensitive data to the cloud, thus increasing security and privacy. Intel's Loihi 2 and Brainchip's Akida processor can be retrained on local data for personalized applications and faster response rates.

This study found the power consumption and performance of processors varies in different architectures, and application domains. For extreme wearable edge devices, the power consumption ranges from 100 μW to a few mW and computethroughput is around 1 GOPS. We found that many applications require higher computing performance, such as video processing, and autonomous car operations. These high-performance applications consume a higher amount of power than extreme edge processors. For example, IBM's NorthPole computes at 200 TOPS with INT8, while consuming 60 W of power. This study found that for the same range of power consumption and chip size, PIM architectures perform better than dataflow or neuromorphic processors. This review found that the PIM processors show significant energy efficiency and consume less power compared to dataflow and neuromorphic processors. For example, the Mythic M1108 is a PIM processor and has the highest performance (35 TOPS), among dataflow and neuromorphic processors that consume less than 10W of power. Neuromorphic processors are highly efficient for performing computation with less synaptic operations but may not be ideal for deep learning applications yet.

There are several emerging deep learning applications that are receiving significant interest. This includes generative AI models, such as transformer models used in ChatGPT and DALL-E for automated art generation. Transformer models are taking the AI world so aggressively, manifested by their super-intelligent chatbot and search queries. Generative AI models are also taking place in image and creative art generation. Transformer engines are mainly designed for data centers or cloud applications, but some processors, such as NVIDIA Hopper H100 [305], can be used for edge workloads. Samsung has released digital PIM for generative AI applications in the data center and edge [306]. ResNet, GoogleNet, and YOLO models are also being used in various industries for facial recognition, lane keeping assistance, and surveillance. Deep reinforcement learning is becoming popular for autonomous learning models in dynamic environments. All of these applications could benefit from highly efficient specialized processors that could run the applications locally, without the need for cloud access. Future directions for industry could be to implement these algorithms in emerging non-Von Neumann computing paradigms for low power computing on edge devices. Current dataflow processors, such as the NVIDIA Orin, or the IBM NorthPole would probably be

 doi:10.20944/preprints202407.0025.v2

able to handle these applications without any changes. More emerging architectures, such as PIM and neuromorphic technologies, may need more enhancements to enable these applications to run on edge devices.

References

1. M. Merenda., 2020. "Edge machine learning for ai-enabled iot devices: A review", *Sensors*, *20*(9), p.2533, March 2020. DOI: 10.3390/s20092533
2. M. P. Vestias et al., "Moving deep learning to the edge", Algorithms, *13*(5), p.125. March 2020. 10.3390/a13050125
3. IBM, "Why organizations are betting on edge computing?", May 2020, Accessed on: June 1, 2023. Available: https://www.ibm.com/thought-leadership/institute-business-value/report/edge-computing.
4. W. Shi et al., "Edge computing: Vision and challenges", *IEEE internet of things journal*, *3*(5), pp.637-646. October 2016. DOI: 10.1109/JIOT.2016.2579198.
5. Statista. "IoT: Number of Connected Devices Worldwide 2015–2025", November 2016, Accessed on: June 5, 2023. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/.
6. J. M. Chabas et al., "New demand, new markets: What edge computing means for hardware companies", *McKinsey & Company, New York, NY, USA, Tech. Rep. November 2018. Accessed on: August 02, 2023.*
7. Google, "Cloud TPU", Available online: https://cloud.google.com/tpu. Accessed on: May 05, 2023.
8. Accenture Lab. "Driving intelligence at the edge with neuromorphic computing", 2021. Accessed on: June 3, 2023. Available: https://www.accenture.com/_acnmedia/PDF-145/Accenture-Neuromorphic-Computing-POV.pdf.
9. Intel Labs. Technology Brief, "Taking /Neuromorphic Computing to the Next Level with Loihi 2", 2021. Accessed on: May 10, 2023. Available: https://www.intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html.
10. F. Akopyan, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip", *IEEE transactions on computer-aided design of integrated circuits and systems*, *34*(10), pp.1537-1557. August 2015. DOI: 10.1109/TCAD.2015.2474396.
11. Videantis. July 2020. Accessed on: June 11, 2022. Available: https://www.videantis.com/videantis-processor-adopted-for-tempo-ai-chip.html.
12. Konikore, 2021. Accessed on: July 10, 2022. "A living Breathing Machine", Available: https://good-design.org/projects/konikore/.
13. Kalray, May 2023. Accessed on: July07, 2023. Available: https://www.kalrayinc.com/press-release/projet-ip-cube/.
14. Brainchip, 2023, Accessed on: June 21, 2023. Available: https://brainchipinc.com/akida-neuromorphic-system-on-chip/.
15. Synsence, May 2023, Accessed on: June 01, 2023. Available: https://www.synsense-neuromorphic.com/technology.
16. Samsung, HBM-PIM, March 2023, Accessed on: July 25, 2023. Available:https://www.samsung.com/semiconductor/solutions/technology/hbm-processing-in-memory/.
17. Upmem, Upmem-PIM, October 2019, Accessed on May 07, 2023. Available online: https://www.upmem.com/nextplatform-com-2019-10-03-accelerating-compute-by-cramming-it-into-dram/.
18. Mythic, 2021. Accessed on: Feb. 05, 2022. Available: https://www.mythic-ai.com/product/m1076-analog-matrix-processor/.
19. Gyrfalcon, Accessed on: March 03, 2023, Available: https://www.gyrfalcontech.ai/solutions/2803s/.
20. Syntiant, January 2021. Accessed on: Feb. 07, 2023. Available: https://www.syntiant.com/ndp101;https://www.syntiant.com/post/the-growing-syntiant-core-family.
21. Leapmind, Efficiera, July 2023. Accessed on: July 06, 2023, Available: https://leapmind.io/en/news/detail/230801/.

 doi:10.20944/preprints202407.0025.v2

22. K. M. Tarwani et al., "Survey on recurrent neural network in natural language processing", *Int. J. Eng. Trends Technol*, *48*, pp.301-304. June 2017. DOI: *10.14445/22315381/IJETT-V48P253.*
23. Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research*, *57*, pp.345-420. October 2015. arXiv:1510.00726v1 [cs.CL].
24. L. Yao et al., "An improved LSTM structure for natural language processing," In *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)* (pp. 565-569). December 2018. DOI: 10.1109/IICSPI.2018.8690387.
25. S. Wang et al., "Learning natural language inference with LSTM," December 2015. *arXiv preprint arXiv:1512.08849.*
26. E. Azari et al., "An Energy-Efficient Reconfigurable LSTM Accelerator for Natural Language Processing," 2019 *IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, December 2019, pp. 4450-4459, DOI: 10.1109/BigData47090.2019.9006030.
27. W. Li et al., "Stance Detection of Microblog Text Based on Two-Channel CNN-GRU Fusion Network," in *IEEE Access*, vol. 7, pp. 145944-145952, Sept. 2019, DOI: 10.1109/ACCESS.2019.2944136.
28. M. Zulqarnain et al., "Efficient processing of GRU based on word embedding for text classification", *JOIV: International Journal on Informatics Visualization*, *3*(4), pp.377-383. Nov 2019. DOI: 10.30630/joiv.3.4.289.
29. Q. Liu et al., "Content-Guided Convolutional Neural Network for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6124-6137, Sept. 2020, DOI: 10.1109/TGRS.2020.2974134.
30. A. Kumar et al., "MobiHisNet: A Lightweight CNN in Mobile Edge Computing for Histopathological Image Classification," in *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17778-17789, Dec.15, 2021, DOI: 10.1109/JIOT.2021.3119520.
31. M. Wang, "Multi-path convolutional neural networks for complex image classification.", Jun. 2015. *arXiv preprint arXiv:1506.04701.*
32. H. Charlton, MacRumors, June 2023. "Apple reportedly planning to switch technology behind A17 bionic chip to cut cost next year", Accessed on: July 05, 2023. Available: https://www.macrumors.com/2023/06/23/apple-to-switch-tech-behind-a17-to-cut-costs/.
33. L. Wang, Taipei Times. "TSMC says new chips to be world's most advanced", May 2023, Accessed on: June 25, 2023. Available: *https://www.taipeitimes.com/News/biz/archives/2023/05/12/2003799625*.
34. Samsung, Exynos, April 2022, Accessed on: February 06, 2023. Available-:https://www.samsung.com/semiconductor/minisite/exynos/products/all-processors/
35. Z. Q. Lin et al., "Edgespeechnets: Highly efficient deep neural networks for speech recognition on the edge." *arXiv preprint*, Nov. 2018. *DOI:*10.48550/arXiv.1810.08559.
36. T. Shen et al., "The analysis of intelligent real-time image recognition technology based on mobile edge computing and deep learning," *Journal of Real-Time Image Processing*, *18*(4), pp.1157-1166. Oct. 2021. DOI: 10.1007/s11554-020-01039-x.
37. P. Subramaniam et al., "Review of security in mobile edge computing with deep learning," In *2019 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 1-5). Mar. 2019. DOI: 10.1109/ICASET.2019.8714349.
38. A. Krizhevsky et. al., "Imagenet classification with deep convolutional neural networks,". *Advances in neural information processing systems*, *25*, pp.1097-1105. Jun. 2017. DOI: 10.1145/3065386.
39. J. Schneible et al., "Anomaly detection on the edge,". In *MILCOM 2017-2017 IEEE Military Communications Conference (MILCOM)* (pp. 678-682). Oct. 2017. DOI: 10.1109/MILCOM.2017.8170817
40. T. Sirojan et al., "Sustainable Deep Learning at Grid Edge for Real-Time High Impedance Fault Detection," in *IEEE Transactions on Sustainable Computing*, vol. 7, no. 2, pp. 346-357, 1 April-June 2022, DOI: 10.1109/TSUSC.2018.2879960.

41. F. Wang et al., "Deep Learning for Edge Computing Applications: A State-of-the-Art Survey," in *IEEE Access*, vol. 8, pp. 58322-58336, 2020, doi: 10.1109/ACCESS.2020.2982411.
42. M.Z. Alom et al., "A State-of-the-Art Survey on Deep Learning Theory and Architectures," *Electronics*. 2019; 8(3):292. Jan. 2019. DOI: 10.3390/electronics8030292.
43. A. Sengupta et al., "Going deeper in spiking neural networks: VGG and residual architectures," *Frontiers in neuroscience*, *13*, p.95. Mar. 2019. DOI: 10.3389/fnins.2019.00095.
44. L Wen et al., "A transfer convolutional neural network for fault diagnosis based on ResNet-50," *Neural Comput & Applic* **32**, 6111–6124 (2020). DOI: 10.1007/s00521-019-04097-w.
45. C. Szegedy *et al*., "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1-9, DOI: 10.1109/CVPR.2015.7298594.
46. DeepVision (Kinara), March 2022. Accessed on: Jan 08, 2023. Available: https://kinara.ai/about-us/.
47. Kneron, Accessed on: Jan 13, 2023. Available: https://www.kneron.com/page/soc/.
48. Q. Wang et al., "N3LDG: A Lightweight Neural Network Library for Natural Language Processing," Beijing *Da Xue Xue Bao* 55, no. 1 (2019): 113-119. Jan. 2019. DOI: 10.13209/j.0479-8023.2018.065.
49. S. Desai et al., "Lightweight convolutional representations for on-device natural language processing," *arXiv preprint*. Feb. 2020. DOI: 10.48550/arXiv.2002.01535.
50. M. Zhang et al., "Libn3l: a lightweight package for neural nlp," In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 225-229. May 2016. DOI: https://aclanthology.org/L16-1034.
51. Y. Tay et al., "Lightweight and efficient neural natural language processing with quaternion networks," arXiv *preprint*, June 2019. *DOI:* 10.48550/arXiv.1906.04393.
52. Gyrfalcon, "LightSpeur 5801S neural accelerator", 2022, Accessed on: December 10, 2022. Available: https://www.gyrfalcontech.ai/solutions/lightspeeur-5801/.
53. D. Liu et al., "Bringing AI to edge: From deep learning's perspective,". Neurocomputing. Volume 485, 7, Pages 297-320, May 2022., :DOI: 10.1016/j.neucom.2021.04.141.
54. H. Li, "Application of IOT deep learning in edge computing: a review,". Academic Journal of Computing & Information Science. 31;4(5). Oct 2021. DOI: 10.25236/AJCIS.2021.040514.
55. S.S. Zaidi et al., "A survey of modern deep learning-based object detection models,". Digital Signal Processing., 103514. Mar 2022. DOI: 10.1016/j.dsp.2022.103514.
56. J. Chen et al., "Deep learning with edge computing: A Review," in *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655-1674, Aug. 2019, DOI: 10.1109/JPROC.2019.2921977.
57. W. Rawat et al., "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," in *Neural Computation*, vol. 29, no. 9, pp. 2352-2449, Sept. 2017, DOI: 10.1162/neco_a_00990.
58. A. A. M. Al-Saffar et al., "Review of deep convolution neural network in image classification", *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, Jakarta, Indonesia, pp. 26-31, Oct. 2017. DOI: 10.1109/ICRAMET.2017.8253139.
59. N. F. Iandola et al., "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and<0.5 MB model size", arXiv *preprint*, Nov 2016. *DOI:* 10.48550/arXiv.1602.07360.
60. A. Elhassouny et al., "Trends in deep convolutional neural Networks architectures: a review", *2019 International Conference of Computer Science and Renewable Energies (ICCSRE)*, Agadir, Morocco, pp. 1-8, Jul. 2019. DOI: 10.1109/ICCSRE.2019.8807741.
61. A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv *preprint*, *Apr. 2017. DOI:* 10.48550/arXiv.1704.04861.

 doi:10.20944/preprints202407.0025.v2

62. M. Sandler et al., "Mobilenetv2: Inverted residuals and linear bottlenecks", ArXiv preprint, Jan 2018. 10.48550/arXiv.1801.04381.
63. A. Howard et al., "Searching for mobilenetv3", In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314-1324. 2019.
64. X. Zhang et al., "Shufflenet: An extremely efficient convolutional neural network for mobile devices", In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848-6856. 2018.
65. M. Ningning et al., "Shufflenet v2: Practical guidelines for efficient cnn architecture design," In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116-131. 2018.
66. T. Mingxing et al., "Efficientnet: Rethinking model scaling for convolutional neural networks," In *International Conference on Machine Learning*, pp. 6105-6114. PMLR, 2019.
67. V. Niv, Hailo blog, "Object detection at the Edge: Making the right choice," AI on the Edge: the Hailo Blog, Oct 2022, Accessd on: Jan 04, 2023. Available on: https://hailo.ai/blog/object-detection-at-the-edge-making-the-right-choice/.
68. Z. -Q. Zhao et al., "Object Detection with deep learning: A review", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212-3232, Nov. 2019, DOI: 10.1109/TNNLS.2018. 2876865..
69. J. Chen and X. Ran, "Deep learning with edge computing: A review," in *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655-1674, Aug. 2019, DOI: 10.1109/JPROC.2019.2921977.
70. J. -M. Hung *et al*., "An 8-Mb DC-Current-Free Binary-to-8b Precision ReRAM Nonvolatile Computing-in-Memory Macro using Time-Space-Readout with 1286.4-21.6TOPS/W for Edge-AI Devices," *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731715.
71. J. Oruh et al., Adegun, "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," in *IEEE Access*, vol. 10, pp. 30069-30079, 2022, DOI: 10.1109/ACCESS.2022.3159339.
72. B. Liu et al., "Time delay recurrent neural network for speech recognition", *Journal of Physics: Conference Series*. Vol. 1229. No. 1. IOP Publishing, 2019. DOI:10.1088/1742-6596/1229/1/012078
73. Y. zhao et al., "The Speechtransformer for Large-scale Mandarin Chinese Speech Recognition," ICASSP *2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 7095-7099, DOI: 10.1109/ICASSP.2019.8682586.
74. J. Oruh et al., "Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition," *IEEE Access*, vol. 10, pp. 30069-30079, 2022, DOI: 10.1109/ACCESS.2022.3159339.
75. M. Omar et al., "Natural Language Processing: Recent Advances, Challenges, and Future Directions," arXiv preprint, 2022 Jan 3, 10.48550/arXiv.2201.00768.
76. Z. Yuan *et al*., "14.2 A 65nm 24.7μJ/Frame 12.3mW Activation-Similarity-Aware Convolutional Neural Network Video Processor Using Hybrid Precision, Inter-Frame Data Reuse and Mixed-Bit-Width Difference-Frame Data Codec," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2020, pp. 232-234, DOI: 10.1109/ISSCC19947.2020.9063155.
77. Geoff Tate, "Advantages of BFloat16 for AI inference, Oct 2019, Accessed on: Jan 07, 2023, Available: https://semiengineering.com/advantages-of-bfloat16-for-ai-inference/.
78. OpenAI, GPT-4: Technical Report, 27 Mar 2023. 10.48550/arXiv.2303.08774.
79. A. Radford et al., "Language models are unsupervised multi-task learners," OpenAI blog. 24;1(8):9. Feb 2019.
80. T. Brown et al., "Language models are few-shot learners", Advances in neural information processing systems. 33:1877-901. 2020.
81. W. Fedus, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity". arXiv preprint, Jan 2021. DOI: 10.48550/arXiv.2101.03961.

82. Q. Cao et al., "DeFormer: Decomposing pre-trained transformers for faster question answering", arXiv preprint, May 2020. DOI: 10.48550/arXiv.2005.00697.
83. Z. Sun et al., "Mobilebert: a compact task-agnostic bert for resource-limited devices," arXiv preprint. 2020 Apr 2020. 10.48550/arXiv.2004.02984
84. D. Garret, "The Synatiant journey and pervasive NDP" Blog Post, Processor, August 2021. Accessed on: May 5, 2022, Availabe: https://www.edge-ai-vision.com/2021/08/the-syntiant-journey-and-the-pervasive-ndp/#:~:text=In%20the%20summer%20of%202019,will%20capitalize%20on%20the%20momentum.
85. NXP, " iMX Application Processors", Accessed on: July 10, 2023, Available:https://www.nxp.com/products/processors-and-microcontrollers/arm-processors/i-mx-applications-processors/i-mx-9-processors:IMX9-PROCESSORS.
86. NXP, "i.MX 8M Plus-Arm Cortex-A53, Machine Learning Vision, Multimedia and Industrial IoT" Accessed on: June 17, 2023, Available:https://www.nxp.com/products/processors-and-microcontrollers/arm-processors/i-mx-applications-processors/i-mx-8-processors/i-mx-8m-plus-arm-cortex-a53-machine-learning-vision-multimedia-and-industrial-iot:IMX8MPLUS.
87. NXP Datasheet, "i.MX 8M Plus SoM datasheet", Accessed on: February 10, 2023. Available: https://www.solid-run.com/wp-content/uploads/2021/06/i.MX8M-Plus-Datasheet-2021-.pdf.
88. A. Deleo, Cision, PRNewwire, "Mythic expands product lineup with new scalable, power-efficient analog matrix processor for edge AI applications", Mythic 1076, Accessed on: May 10, 2023. Available: https://www.prnewswire.com/news-releases/mythic-expands-product-lineup-with-new-scalable-power-efficient-analog-matrix-processor-for-edge-ai-applications-301306344.html.
89. S. W. Foxton, EETimes. "Mythic Launches Second AI Chip", Accessed on: April 20, 2022. Available: https://www.eetasia.com/mythic-launches-second-ai-chip/.
90. L. Fick et al., "Analog Matrix Processor for Edge AI Real-Time Video Analytics," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, 2022, pp. 260-262.
91. Gyrfalcon, "PIM AI Accelerators", Accessed on: August 01, 2023. Available: https://www.gyrfalcontech.ai/.
92. Gyrfalcon Technology, "Lightspeeur 2803 Neural Accelerator", Accessed on: August 02, 2023. Available: https://www.gyrfalcontech.ai/solutions/2803s/.
93. C. Yu et al., "A survey of model compression and acceleration for deep neural networks." *arXiv preprint, June 2020. DOI:* 10.48550/arXiv.1710.09282.
94. L. Deng et al., "Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey," in *Proceedings of the IEEE*, vol. 108, no. 4, pp. 485-532, April 2020, DOI: 10.1109/JPROC.2020.2976475.
95. K. Nan et al., "Deep model compression for mobile platforms: A survey," in *Tsinghua Science and Technology*, vol. 24, no. 6, pp. 677-693, Dec. 2019, DOI: 10.26599/TST.2018.9010103.
96. A. Berthelier et al., "Deep Model Compression and Architecture Optimization for Embedded Systems: A Survey". *J Sign Process Syst* **93**, 863–878. August 2021. DOI:10.1007/s11265-020-01596-1.
97. J. Lei et al., "A Review of Deep Network Model Compression", Journal of Software, 2018, 29(2): 251-266. DOI: http://www.jos.org.cn/1000- 9825/5428.htm.
98. S. Han et al., "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." *arXiv preprint, DOI:* 10.48550/arXiv.1510.00149.

99. Q. Qin et al., "To compress, or not to compress: Characterizing deep learning model compression for embedded inference." In *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, pp. 729-736. IEEE, 2018. DOI: 10.1109/BDCloud.2018.00110.
100. B. Jacob *et al.*, "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 2704-2713, DOI: 10.1109/CVPR.2018.00286.
101. Y. Chunyu, and S. S. Agaian, "A comprehensive review of Binary Neural Network." *arXiv preprint, Mar. 2023. DOI:* 10.1007/s10462-023-10464-w.
102. H. Mo *et al.*, "9.2 A 28nm 12.1TOPS/W Dual-Mode CNN Processor Using Effective-Weight-Based Convolution and Error-Compensation-Based Prediction," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, 2021, pp. 146-148. DOI: 10.1109/ISSCC42613.2021.9365943.
103. S. Yin *et al.*, "PIMCA: A 3.4-Mb Programmable In-Memory Computing Accelerator in 28nm for On-Chip DNN Inference," *2021 Symposium on VLSI Circuits*, 2021, pp. 1-2. DOI: 10.23919/VLSICircuits52068.2021.9492403.
104. H. Fujiwara *et al.*, "A 5-nm 254-TOPS/W 221-TOPS/mm2 Fully Digital Computing-in-Memory Macro Supporting Wide-Range Dynamic-Voltage-Frequency Scaling and Simultaneous MAC and Write Operations," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731754.
105. S. Wang and P. Kanwar, "BFloat16: The secret to high performance on Cloud TPUs", Aug. 2019. Accessed on: Sept. 18, 2022, Available: https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus.
106. G. Tate, "Advantages of BFloat16 for AI inference, Oct 2019, Accessed on: Sept. 18, 2022. Available: https://semiengineering.com/advantages-of-bfloat16-for-ai-inference/.
107. S. Lee *et al.*, "A 1ynm 1.25V 8Gb, 16Gb/s/pin GDDR6-based Accelerator-in-Memory supporting 1TFLOPS MAC Operation and Various Activation Functions for Deep-Learning Applications," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, doi: 10.1109/ISSCC42614.2022.9731711.
108. Blaize, Stop Compromising Start Deploying, 2022, Accessed on: Jun 11, 2023. Available: https://www.blaize.com/products/ai-edge-computing-platforms/.
109. M. Demer, Blaize Ignites Edge-AI Performance, Microprocessor Report, Sept. 2020. Accessed on: Jun. 2022. Available: https://www.blaize.com/wp-content/uploads/2020/09/Blaize-Ignites-Edge-AI-Performance.pdf.
110. T. Liang et al., "Pruning and quantization for deep neural network acceleration: A survey." *Neurocomputing* 461 (2021): 370-403. Oct. 2021. DOI: 10.1016/j.neucom.2021.07.045.
111. B. M. Mahdi, and M. Ghatee. "A systematic review on overfitting control in shallow and deep neural networks." *Artificial Intelligence Review*, 1-48. *Dec.* 2021, DOI: 0.1007/s10462-021-09975-1
112. H. Yang et al., "Soft filter pruning for accelerating deep convolutional neural networks." *arXiv preprint. Aug. 2018. DOI:* 10.48550/arXiv.1808.06866.
113. H. Torsten et al., "Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks." *arXiv preprint, Jan.2021.DOI:* 10.48550/arXiv.2102.00554.
114. V. Sanh eta al., "Movement pruning: Adaptive sparsity by fine-tuning." *arXiv preprint, Oct. 2020. DOI:* 10.48550/arXiv.2005.07683.

115. B. Cristian et al. "Model compression." In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535-541. Aug. 2006. DOI: 10.1145/1150402.1150464.
116. G. Jianping et al., "Knowledge distillation: A survey." *International Journal of Computer Vision* 129, no. 6: 1789-1819. May 2021. DOI: 10.48550/arXiv.2006.05525.
117. Kim Yoon, and Alexander M. Rush. "Sequence-level knowledge distillation." *arXiv preprint, Sep. 2016, DOI:* 10.48550/arXiv.1606.07947.
118. A. -Z, Zeyuan, and Y. Li. "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning." *arXiv preprint, Feb 2023,* 10.48550/arXiv.2012.09816.
119. M. Huang et al., Knowledge Distillation for Sequence Model. In *Interspeech* (pp. 3703-3707). Sep 2018. DOI: 10.21437/Interspeech.2018-1589.
120. C. J. Hyun, and B. Hariharan. "On the efficacy of knowledge distillation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4794-4802. Oct 2019.
121. T. Tambe et al., "EdgeBERT: Optimizing On-chip inference for multi-task NLP", arXiv preprints. Nov 2020. DOI: 10.48550/arXiv.2011.14203.
122. Z. Sun et al., "Mobilebert: a compact task-agnostic bert for resource-limited devices", arXiv preprint Apr 6, 2020. DOI: 10.48550/arXiv.2004.02984.
123. Tensorflow, "An end-to-end open-source machine learning platform", Accessed on: May 01, 2023. Available: https://www.tensorflow.org/.
124. S. Li, "TensorFlow Lite: On-Device Machine Learning Framework", *Journal of Computer Research and Development*, 2020, 57(9): 1839-1853. DOI: 10.7544/issn1000-1239.2020.20200291.
125. A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library", *Advances in neural information processing systems*, *32*, pp.8026-8037. Dec 2019.
126. Pytorch, Pytorch Mobile, "End to end workflow from training to deployment for iOS and android mobile devices", Accessed on: Dec 20, 2022. Available: https://pytorch.org/mobile/home/.
127. Keras, "Keras API References", Accessed on: Dec 20, 2022. Available: Online link: https://keras.io/api/.
128. Caffe2, "A new lightweight, modular, and scalable deep learning framework", Accessed on: Dec 21, 2022. Available: https://research.facebook.com/downloads/caffe2/.
129. A. Zelinsky, "Learning OpenCV---Computer Vision with the OpenCV Library (Bradski, G.R. et al.; 2008) [On the Shelf]," in *IEEE Robotics & Automation Magazine*, vol. 16, no. 3, pp. 100-100, September 2009, DOI: 10.1109/MRA.2009.933612.
130. ONNX, "Open Neural Network Exchange-the open standard for machine learning interoperability", Accessed on: Dec 22, 2022. Available: https://onnx.ai/.
131. MXNet, "A flexible and efficient efficient library for deep learning", Accessed on: Dec 22, 2022. Available: https://mxnet.apache.org/versions/1.9.0/.
132. ONNX, "Meta AI", Accessed on: Dec 23, 2022, Available: https://ai.facebook.com/tools/onnx/.
133. P. Vajda, and Y. Jia, "Delivering real-time AI in the palm of your hand", Accessed on: Dec 27, 2022. Available: https://engineering.fb.com/2016/11/08/android/delivering-real-time-ai-in-the-palm-of-your-hand/.
134. CEVA, "Edge AI & deep learning", Accessed on: July 10, 2023. Available: https://www.ceva-dsp.com/app/deep-learning/.
135. M. Demler, "CEVA Neupro Accelerator Neural Nets", Microprocessor Report, Jan 2018. Available:https://www.ceva-dsp.com/wp-content/uploads/2018/02/Ceva-NeuPro-Accelerates-Neural-Nets.pdf.
136. CEVA, "CEVA NeuPro-S On-device Computer Vision Processor Architecture", Sep 2020, Accessed on: Jun 17, 2022. Available: https://www.ceva-dsp.com/wp content/uploads/2020/11/09_11_20_NeuPro-S_Brochure_V2.pdf.

137. P.A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface". Science. 8;345(6197):668-73. Aug 2014. DOI: 10.1126/science.1254642.
138. C. Yakopcic et al., "Solving Constraint Satisfaction Problems Using the Loihi Spiking Neuromorphic Processor," *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Grenoble, France, 2020, pp. 1079-1084, doi: 10.23919/DATE48585.2020.9116227.
139. T. Bohnstingl, "Neuromorphic hardware learns to learn", Frontiers in neuroscience, 21;13:483. May 2019. DOI: 10.3389/fnins.2019.00483
140. S.B. Shrestha et al., "Slayer: Spike layer error reassignment in time", Advances in neural information processing systems,31. Sep 2018. DOI: 10.48550/arXiv.1810.08646.
141. S. Davidson, S. B. Furber, "Comparison of artificial and spiking neural networks on digital hardware", Frontiers in Neuroscience, 15:345. Apr 2021. DOI: 10.3389/fnins.2021.651141.
142. P. Blouw et al., "Benchmarking keyword spotting efficiency on neuromorphic hardware", In Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop, pp. 1-8. Mar 2019, DOI: 10.1145/3320288.3320304.
143. NengoLoihi, Accessed on: Nov 20, 2022. Available: https://www.nengo.ai/nengo-loihi/.
144. Nengo, "Spinnaker backend for Nengo", Accessed on: Nov 20, 2022. Available: https://nengo-spinnaker.readthedocs.io/en/latest/.
145. NengoDL, Accessed on: Nov 20, 2022, Available: https://www.nengo.ai/nengo-dl/.
146. Brainchip, MetaTF, Online link: https://brainchip.com/metatf-development-environment/.
147. Brainchip, "Introducing the ADK1000 IP and NSOM for Edge AI IoT", May 2020, Accessed on: Nov 22, 2022. Available: https://www.youtube.com/watch?v=EUGx45BCKIE.
148. P. Clarke, eeNews, "Akida Spiking Neural Processor Could Head to FDSOI", Aug 2, 2021, Accessed on: Nov 25, 2022. Available: https://www.eenewsanalog.com/news/akida-spiking-neural-processor-could-head-fdsoi.
149. M. Demer, "Brainchip Akida is a faster learner", microprocessor report, Lynely group, Oct 28, 2019. Available: https://d1io3yog0oux5.cloudfront.net/brainchipinc/files/BrainChip+Akida+Is+a+Fast+Learner.pdf.
150. Lava. "Lava software framework", Accessed on: Nov 26, 2022. Available: https://lava-nc.org/.
151. A. Reuther et al., "AI and ML Accelerator Survey and Trends," *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, Waltham, MA, USA, 2022, pp. 1-10, doi: 10.1109/HPEC55821.2022.9926331.
152. Y. Chen et al., "A survey of accelerator architectures for deep neural networks", *Engineering*, *6*(3), pp.264-274. Mar 2020. DOI: 10.1016/j.eng.2020.01.007.
153. W. Li and M. Liewig, "A survey of AI accelerators for edge environments", In *World Conference on Information Systems and Technologies* (pp. 35-44). Springer, Cham. April 2020. DOI: 10.1007/978-3-030-45691-7_4.
154. M. S. Murshed et al., "Machine learning at the network edge: A survey", ACM Computing Surveys (CSUR). 54(8):1-37. Oct 2021, DOI: 10.1145/3469029.
155. W. Lin et al., "Low-Power Ultra-Small Edge AI Accelerators for Image Recognition with Convolution Neural Networks: Analysis and Future Directions", Electronics, *10*(17), p.2048. Aug 2021. DOI: 10.3390/electronics10172048.
156. A. Reuther et al., "Survey of Machine Learning Accelerators," *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, Waltham, MA, USA, 2020, pp. 1-12, doi: 10.1109/HPEC43674.2020.9286149.
157. J. Cross. Macworld. "Apple's A16 chip doesn't live up to its 'Pro' price or expectations", Accessed on: Jan 1, 2023. Available: https://www.macworld.com/article/1073243/a16-processor-cpu-gpu-lpddr5-memory-performance.html.

158. Apple. Press Release, June 6, 2022. "Apple unveils M2, taking the breakthrough performance and capabilities of M1 even further", Accessed on: July 10, 2022. Available: https://www.apple.com/newsroom/2022/06/apple-unveils-m2-with-breakthrough-performance-and-capabilities/.
159. Apple. Press Release. Nov 10, 2020. "Apple Unleashes M1", Accessed on: Dec 5, 2021. Available: https://www.apple.com/newsroom/2020/11/apple-unleashes-m1/.
160. ARM, NPU, Ethos-78, "Highly scalaeable and efficient second generation ML inference processor", Accessed on: May 15, 2022. Available: https://www.arm.com/products/silicon-ip-cpu/ethos/ethos-n78.
161. A. Frumusanu, "Arm Announces Ethos-N78: Bigger and More Efficient", Anandtech, May 27, 2020. Accessed on: April 25, 2022. Available: https://www.anandtech.com/show/15817/arm-announces-ethosn78-npu-bigger-and-more-efficient.
162. Blaize. "2022 best edge AI processor blaize Pathfinder P1600 embedded system on module", Accessed on: Dec 05, 2022. Available: https://www.blaize.com/products/ai-edge-computing-platforms/.
163. M. Demer, "Blaize Ignites Edge-AI Performance", Microprocessor Report, Sep 2020, Accessed on: June 20, 2022. Available: https://www.blaize.com/wp-content/uploads/2020/09/Blaize-Ignites-Edge-AI-Performance.pdf.
164. AIMotive. "Industry High 98% Efficiency Demonstrated Aimotive and Nextchip", April 15, 2021, Accessed on: Mar 25, 2022. Available: https://aimotive.com/-/industry-high-98-efficiency-demonstrated-by-aimotive-and-nextchip.
165. AIMotive. "NN acceleration for automotive AI", Accessed on: May 25, 2022. Available: https://aimotive.com/aiware-apache5.
166. N. Dahad, "Hardware Inference Chip Targets Automotive Applications", December 24, 201. Accessed on: June 25, 2022. Available:https://www.embedded.com/hardware-inference-chip-targets-automotive-applications/..
167. Cadence, "Tesilica AI Platform", Accessed on: Dec 12, 2022. Available: https://www.cadence.com/en_US/home/tools/ip/tensilica-ip/tensilica-ai-platform.html.
168. Cadence Newsroom, "Cadence Accelerates Intelligent SoC Development with Comprehensive On-Device Tensilica AI Platform", Sep 13, 2021. Accessed on: Aug 15, 2022. Available: https://www.cadence.com/en_US/home/company/newsroom/press-releases/pr/2021/cadence-accelerates-intelligent-soc-development-with-comprehensi.html.
169. M. Maxfield, "Say Hello to Deep Vision's Polymorphic Dataflow Architecture", EE Journal, Dec 24, 2020. Accessed on: Dec05, 2022, Available: https://www.eejournal.com/article/say-hello-to-deep-visions-polymorphic-dataflow-architecture/.
170. S. Ward-Foxton, "AI Startup Deepvision Raises Funds Preps Next Chip", EETimes, Sep 15, 2021, Accessed on: Dec 05, 2022. Available: https://www.eetasia.com/ai-startup-deep-vision-raises-funds-preps-next-chip/.
171. Horizon AI, "Efficient AI Computing for Automotive Intelligence", Accessed on: Dec 06, 2022, Available: https://en.horizon.ai/.
172. Horizon Robotics, "Horizon Robotics and BYD announce cooperation on BYD's BEV perception solution powered by Journey 5 computing solution at Shanghai auton show 2023," Cision PR Newswire, Apr 19, 2023. Accessed on: Jun 20, 2023. Available: https://www.prnewswire.com/news-releases/horizon-robotics-and-byd-announce-cooperation-on-byds-bev-perception-solution-powered-by-journey-5-computing-solution-at-shanghai-auto-show-2023-301802072.html.
173. D. Zheng, "Horizon Robotics' AI chip with up to 128 TOPS computing power gets key certification", Cnevpost, Jul 6, 2021. Accessed on June 16, 2022. Available: https://cnevpost.com/2021/07/06/horizon-robotics-ai-chip-with-up-to-128-tops-computing-power-gets-key-certification/.
174. Hailo, "The World's Top Performance AI Processor for Edge Devices", Accessed on: May 20, 2023. Available: https://hailo.ai/.

175. E. Brown, "Hailo-8 NPU Ships on Linux-Powered Lanner Edge System", Jun 1, 2021. Accessed on: Jul 10, 2022. Available: https://linuxgizmos.com/hailo-8-npu-ships-on-linux-powered-lanner-edge-systems/.
176. Edge TPU. "Coral Technology", Accessed on: May 20, 2022. Available: https://coral.ai/technology/
177. Coral, "USB Accelerator", Accessed on: Jun 13, 2022. Available: https://coral.ai/products/accelerator/.
178. N. P. Jouppi et al., "A domain-specific supercomputer for training deep neural networks", Communications of the ACM.;63(7):67-78. Jun 2020. DOI: 10.1145/3360307
179. Google, "How Google Tensor powers up Pixel phones", Accessed on: Jul 16, 2022. Available: https://store.google.com/intl/en/ideas/articles/google-tensor-pixel-smartphone/.
180. GreenWaves, "GAP9 Processor for Hearables and Sensors", Accessed on: Jun 18, 2023. Available: https://greenwaves-technologies.com/gap9_processor/.
181. A. Deleo, GreenWaves. GAP9, "GreenWaves Unveils Groundbreaking Ultra-Low Power GAP9 IoT Application Processor for the Next Wave of Intelligence at the Very Edge", Accessed on: Aug 08, 2023. Avilable: https://greenwaves-technologies.com/gap9_iot_application_processor/.
182. Imagination, "Power Series3NX, Advanced Compute and Neural Network Processors Enabling the Smart Edge", Accesed on: Jun 10, 2022. Available: https://www.imaginationtech.com/vision-ai/powervr-series3nx/.
183. B. Har-Evan, "Seperating the wheat from the chaff in embedded AI with PowerVR Series3NX", Jan 24, 2019. Accessed on: Jul 25, 2022. Available: https://www.imaginationtech.com/blog/separating-the-wheat-from-the-chaff-in-embedded-ai/.
184. Imagination, "The ideal single core solution for neural network acceleration", Accessed on: June 16, 2022. Available: https://www.imaginationtech.com/product/img-4nx-mc1/.
185. Wikichip, "Intel Nirvana, Neural Network Processor (NNP)", Accessed on: July 14, 2023, Available: https://en.wikichip.org/wiki/nervana/nnp.
186. Carmelito, "Intel Neural Compute Stick 2-Review", Element14, Mar 8, 2021. Accessed on: Mar 24, 2023. Availabel: https://community.element14.com/products/roadtest/rv/roadtest_reviews/954/intel_neural_compute_3.
187. L. Smith, "4th Gen Intel Xeon Scalable Processors Launched", StorageReview, Jan 10, 2023. Accessed on: May 12, 2023. Available: https://www.storagereview.com/news/4th-gen-intel-xeon-scalable-processors-launched.
188. J. Burns, and L. Chang, "Meet the IBM Artificial Intelligence Unit", Oct 18, 2022. Accessed on: Dec 16, 2022. Available: https://research.ibm.com/blog/ibm-artificial-intelligence-unit-aiu
189. K. Gupta. "IBM Research Introduces Artificial Intelligence Unit (AIU): It's First Complete System-on-Chip Designed to Run and Train Deep Learning Models Faster and More Efficiently than a General-Purpose CPU", MarkTecPost, Oct 27, 2022. Accessed on: Dec 20, 2022. Available: https://www.marktechpost.com/2022/10/27/ibm-research-introduces-artificial-intelligence-unit-aiu-its-first-complete-system-on-chip-designed-to-run-and-train-deep-learning-models-faster-and-more-efficiently-than-a-general-purpose-cpu/.
190. P. Clarke, "Startup launches near-binary neural network accelerator", EENews, May 19, 2020. Accessed on: Dec 20, 2022. Available: https://www.eenewseurope.com/en/startup-launches-near-binary-neural-network-accelerator/.
191. Kneron, "AI System on Chip (SoC)", KL720 AI SoC, Accessed on: Jul 15, 2023. Available: https://www.kneron.com/page/soc/.
192. Kneron, ""AI System on Chip (SoC)", Accessed on: Jul 15, 2023", KL530 AI SoC, https://www.kneron.com/page/soc/ .
193. MobileEye, "One automatic grade SoC, many mobility solutions", Accessed on Aug 4, 2023. Available: https://www.mobileye.com/our-technology/evolution-eyeq-chip/.
194. EyeQ5, Wikichip, March 2021. Accessed on: Jun 22, 2023. Availabe: https://en.wikichip.org/wiki/mobileye/eyeq/eyeq5.

 doi:10.20944/preprints202407.0025.v2

195. D. Casil, "Mobileye presents EyeQ Ultra, the chip that promises true level 4 autonomous driving in 2025", Jul 01, 2022. Accessed on: Jun 05, 2023. Available: https://www.gearrice.com/update/mobileye-presents-eyeq-ultra-the-chip-that-promises-true-level-4-autonomous-driving-in-2025/.
196. MobileEye, "Meet EyeQ6: Our most advanced driver-assistance chip yet", May 25, 2022. Accessed on: May 27, 2023. Available: https://www.mobileye.com/blog/eyeq6-system-on-chip/.
197. Nvidia, "Jetson Nano", Accessed on: May 26, 2023. Available: https://elinux.org/Jetson_Nano#:~:text=Useful%20for%20deploying%20computer%20vision,5%2D10W%20of%20power%20consumption.
198. NIDIA Jetson Nano B01, "Deep learning with raspberry pi and alternatives" April 5, 2023. Accessed on Jul 03, 2023. Available: https://qengineering.eu/deep-learning-with-raspberry-pi-and-alternatives.html#Compare_Jetson.
199. Nvidia, Jetson Orin, "The future of industrial-grade edge AI", Accessed on: Jul 25, 2023. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/.
200. L. Deleon, "Build enhanced video conference experiences", Qualcom. Mar 7, 2023. Accessed on: May 05, 2023, Available: https://developer.qualcomm.com/blog/build-enhanced-video-conference-experiences.
201. Qualcomm, QCS8250, "Premium processor designed to help you deliver maximum performance for compute intensive camera, video conferencing and Edge AI applications with support Wi-Fi 6 and 5G for the Internet of Things (IoT)" Accessed on: Jul, 15, 2023, https://www.qualcomm.com/products/qcs8250.
202. Snapdragon, "888+ 5G Mobile Platform", Accessed on: May 24, 2023. Available: https://www.qualcomm.com/products/snapdragon-888-plus-5g-mobile-platform.
203. Qualcomm, "Qualcomm Snapdragon 888 Plus, Benchmark, Test and Spec", CPU monkey, Jun 16, 2023, Accessed on: Jul 15, 2023. Available: https://www.cpu-monkey.com/en/cpu-qualcomm_snapdragon_888_plus.
204. H. Hsu, "Training ML Models at the Edge with Federated Learning", Qualcomm, Jun 07, 2021. Accessed on: Jul 7, 2023. Available: https://developer.qualcomm.com/blog/training-ml-models-edge-federated-learning.
205. Samsung, "The core that redefines your device", Accessed on May 25, 2023.Available:https://www.samsung.com/semiconductor/minisite/exynos/products/all-processors/.
206. GSMARENA, "Exynos 2100 Vs Snapdragon 888: Benchmarking the Samsung Galaxy S21 Ultra Versions", GSMARENA, Feb 07, 2021. Accessed on: Jun 10, 2023.Available:https://www.gsmarena.com/exynos_2100_vs_snapdragon_888_benchmarking_the_samsung_galaxy_s21_ultra_performance-news-47611.php.
207. M. Kong, "VeriSilicon VIP9000 NPU AI processor and ZSPNano DSP IP bring AI-Vision and AI-Voice to low power automotive image processing SoC", VeriSilicon Press release, May 12, 2020. Acessed on: Jul 16, 2022. Available:https://www.verisilicon.com/en/PressRelease/VIP9000andZSPAdoptedbyiCatch.
208. VeriSilicon, "VeriSilicon Launches VIP9000, New Generation of Neural Processor Unit IP", VeriSilicon Press Release, Jul 8, 2019. Accessed on: May 25, 2022. Available: https://www.verisilicon.com/en/PressRelease/VIP9000.
209. Synopsys, "Designware ARC EV Processors for Embedded Vsion", Accessed on: Jul 25, 2022. Available: https://www.synopsys.com/designware-ip/processor-solutions/ev-processors.html.
210. Synopsys, "Synopsys EV7x vision processor", Accessed on: May 25, 2023. Available:https://www.synopsys.com/dw/ipdir.php?ds=ev7x-vision-processors.
211. Wikichip, "FSD Chip", Wikichip, Accessed on: May 28, 2023. Available: https://en.wikichip.org/wiki/tesla_(car_company)/fsd_chip.
212. Research and Markets, "Neuromorphic Chips: Global Strategic Business Report", Research and Markets, . ID: 4805280, Accessed on: May 16, 2023, Available:https://www.researchandmarkets.com/reports/4805280/neuromorphic-chips-global-strategic-business.

213. M. Ghilardi, "Synsense secures additional capital from strategic investors", News Synsecse, Apr 18, 2023. Accessed on: May 5, 2023. Available: https://www.venturelab.swiss/SynSense-secures-additional-capital-from-strategic-investors.
214. GrAI VIP, "Life Ready AI Processors", Accessed on: Jul 16, 2023. Available: https://www.graimatterlabs.ai/product.
215. A. S. Cassidy *et al*., "Real-Time Scalable Cortical Computing at 46 Giga-Synaptic OPS/Watt with ~100× Speedup in Time-to-Solution and ~100,000× Reduction in Energy-to-Solution," *SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, New Orleans, LA, USA, 2014, pp. 27-38, DOI: 10.1109/SC.2014.8.
216. S. Wax-forton, "Innatera unveils neuromorphic AI chip to accelerate spiking networks", EETimes, Jul 7, 2021. Accessed on: May 25, 2023, Available:https://www.linleygroup.com/newsletters/newsletter_detail.php?num=6302&year=2021&tag=3.
217. J L Aufrace, "Innatera Neuromorphic AI Accelerator for Spiking Neural Networks Enables Sub-mW AI Inference", CNX software-embedded systems news, Jul 16, 2021. Accessed on: May 25, 2023. Available: https://www.cnx-software.com/2021/07/16/innatera-neuromorphic-ai-accelerator-for-spiking-neural-networks-snn-enables-sub-mw-ai-inference/.
218. B. Rajendran et al., "Low-Power Neuromorphic Hardware for Signal Processing Applications: A Review of Architectural and System-Level Design Approaches," in *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 97-110, Nov. 2019, DOI: 10.1109/MSP.2019.2933719.
219. Blouw P, Choo X, Hunsberger E, Eliasmith C. Benchmarking keyword spotting efficiency on neuromorphic hardware. In Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop 2019 Mar 26 (pp. 1-8).
220. Yousefzadeh A et al., SENeCA: Scalable energy-efficient neuromorphic computer architecture. In2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS) 2022 Jun 13 (pp. 371-374). IEEE
221. Konikore, "Technology that sniffs out danger", Accessed on: May 26, 2023. Available: https://theindexproject.org/post/konikore.
222. K. Ueyoshi *et al*., "DIANA: An End-to-End Energy-Efficient Digital and ANAlog Hybrid Neural Network SoC," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, 2022, pp. 1-3, doi: 10.1109/ISSCC42614.2022.9731716.
223. N. Flaherty, "Axelera shows DIANA analog in-memory computing chip", EENews, Feb 21, 2022. Accessed on: Jul 22, 2023. Available: https://www.eenewseurope.com/en/axelera-shows-diana-analog-in-memory-computing-chip/.
224. Gyrfalcon Technology, "PIM AI Accelerators", Accessed on: Mar 25, 2022. Available: https://www.gyrfalcontech.ai/.
225. SolidRun, "Janux GS31 AI Server", Accessed on: Mar 25, 2022. Available: https://www.solid-run.com/embedded-networking/nxp-lx2160a-family/ai-inference-server/.
226. Samsung, "Samsung Brings PIM Technology to Wider Applications", Aug 24, 2021. Accessed on: May 18. 2023. Available: https://www.samsung.com/semiconductor/newsroom/news-events/samsung-brings-in-memory-processing-power-to-wider-range-of-applications/.
227. Kim JH, Kang SH, Lee S, Kim H, Song W, Ro Y, Lee S, Wang D, Shin H, Phuah B, Choi J. Aquabolt-XL: Samsung HBM2-PIM with in-memory processing for ML accelerators and beyond. In2021 IEEE Hot Chips 33 Symposium (HCS) 2021 Aug 22 (pp. 1-26). IEEE.
228. Syntiant, "Making Edge AI a Reality: A new processor for deep learning", Accessed on: Jun 28, 2023. Availabel: https://www.syntiant.com/.

229. Syntiant, "NDP100 Neural Decision Processor- NDP100- always-on speech recognition", Accessed on: Jun 28, 2023. Available: https://www.syntiant.com/ndp100.
230. Syntiant, "NDP200 Neural Decision Processor, NDP200 always-on vision, sensor and speech recognition", Accessed on: Jun 28, 2023. Available: https://www.syntiant.com/ndp200.
231. N. Tyler, "Syntiant Introduces NDP102 Neural Decision Processor", newelectronics, Sep 16, 2021. Accessed on: Jun 28, 2023. Aailable: https://www.newelectronics.co.uk/content/news/syntiant-introduces-ndp102-neural-decision-processor.
232. G. Halfacree, "Syntiant's NDP200 Promises 6.4GOP/s of Edge AI Compute in a Tiny 1mW Power Envelope", Hackster.io, 2021, Accessed on: Jun 29, 2023. Available:https://www.hackster.io/news/syntiant-s-ndp200-promises-6-4gop-s-of-edge-ai-compute-in-a-tiny-1mw-power-envelope-96590283ffbc.
233. M. Demler, "Syntiant Knows All the Best Words, NDP10x Speech-Recognition Processors Consume Just 200uW", Microprocessors Report, 2019, Accessed on: Jun 29, 2023. Available: https://www.syntiant.com/post/syntiant-knows-all-the-best-words.
234. M. Demler, "Syntiant NDP120 Sharpens Its Hearing, Wake-Word Detector COmbines Ultra-Low Power DLA with HiFi 3DSP", 2021, Available: https://www.linleygroup.com/mpr/article.php?id=12455.
235. G. Medici, "Syntiant Introduces NDP102 Neural Decision Processor", Syntiant, Sep 15, 2021. Accessed on: Jun 30, 2023. Available: https://www.newelectronics.co.uk/content/news/syntiant-introduces-ndp102-neural-decision-processor.
236. Untether, "The most efficient AI computer engine available", Accessed on: May 18, 2023. Available: https://www.untether.ai/press-releases/untether-ai-ushers-in-the-petaops-era-with-at-memory-computation-for-ai-inference-workloads.
237. Untether, "Untether AI", Accessed on: May 18, 2023. Available: https://www.colfax-intl.com/downloads/UntetherAI-tsunAImi-Product-Brief.pdf.
238. Upmem, "The PIM reference platform", Accessed on: May 19, 2023. Available:https://www.upmem.com/technology/.
239. D. Lavenier, R. Cimadomo and R. Jodin, "Variant Calling Parallelization on Processor-in-Memory Architecture," *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Seoul, Korea (South), 2020, pp. 204-207, DOI: 10.1109/BIBM49941.2020.9313351.
240. J. Gómez-Luna et al., "Benchmarking Memory-Centric Computing Systems: Analysis of Real Processing-in-Memory Hardware", *arXiv preprint, DOI:* 10.48550/arXiv.2110.01709.
241. Ian Cutress, "Hot Chips 31 Analysis: In Memory Processing by Upmem", Anandtech, Aug 18, 2019. Accessed on: May 20, 2023. Available: https://www.anandtech.com/show/14750/hot-chips-31-analysis-inmemory-processing-by-upmem.
242. Nanoreview.net., "A14 Bionic vs. A15 Bionic", Accessed on: Jun 16, 2023. Available:https://nanoreview.net/en/soc-compare/apple-a15-bionic-vs-apple-a14-bionic.
243. R. Merrit, "Startup Accelerates AI at the Sensor", EETimes, Feb 11, 2019. Accessed on: Jun 10, 2023. Available:https://www.eetimes.com/startup-accelerates-ai-at-the-sensor/.
244. P. Clarke, "Indo-US Startup Preps Agent-based AI Processor", EENews, Aug 26, 2018. Accessed on: Jun 20, 2023. Available: https://www.eenewsanalog.com/en/indo-us-startup-preps-agent-based-ai-processor-2/.
245. B. Wheeler, "Bitmain SoC Brings AI to the Edge", Accessed on: Jul 23, 2023. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php?num=5975&year=2019&tag=3.
246. W. Liang, "Get Started, Neural Network Stick", Github, May 10, 2019. Accessed on: May 16, 2023. Available: https://github.com/BM1880-BIRD/bm1880-system-sdk/wiki/GET-STARTED.

 doi:10.20944/preprints202407.0025.v2

247. L. Gwennap, "Kendryte Embeds AI for Surveillance", Accessed on: Jul 14, 2023. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php?num=5992.
248. Canaan, "Kendryte K210", Accessed on: May 15, 2023. Available: https://canaan.io/product/kendryteai.
249. Eta Compute, "Micropower AI vision platform", Accessed on: May 15, 2023. Available: https://etacompute.com/tensai-flow/.
250. FlexLogic, "Flexlogic announces InferX high performance IP for DSP and AI inference" Apr 24, 2023. Accessed on Jun 12, 2023, Available: https://flex-logix.com/inferx-ai/inferx-ai-hardware/. 2023.
251. Mediatek, i350, "Mediatek introduces i350 edge AI platform designed for voice and vision processing applications", Oct 14, 2020, Accessed on: May 16, 2023. Available: https://corp.mediatek.com/news-events/press-releases/mediatek-introduces-i350-edge-ai-platform-designed-for-voice-and-vision-processing-applications.
252. Perceive, "Put high power intelligence in a low poer device", Accessed on: May 16, 2023. Available: https://perceive.io/product/ergo/.
253. Yida, "Introducing the Rock Pi N10 RK3399Pro SBC for AI and Deep Learning", Accessed on: May 17, 2023. Available: https://www.seeedstudio.com/blog/2019/12/04/introducing-the-rock-pi-n10-rk3399pro-sbc-for-ai-and-deep-learning/.
254. GadgetVersus, "Amalogic A311D processor benchmarks and Specs", Accessed on: May 16, 2023. Available: https://gadgetversus.com/processor/amlogic-a311d-specs/.
255. Samsung, "Exynos 2200", Accessed on: Jun 1, 2023. Available: https://semiconductor.samsung.com/us/processor/mobile-processor/exynos-2200/.
256. Think Silicon, "Nema Pico XS", Accessed on: May 23, 2023. Available: https://www.think-silicon.com/nema-pico-xs#features.
257. Y. -D. Chih *et al.*, "16.4 An 89TOPS/W and 16.3TOPS/mm2 All-Digital SRAM-Based Full-Precision Compute-In Memory Macro in 22nm for Machine-Learning Edge Applications," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 252-254, DOI: 10.1109/ISSCC42613.2021.9365766.
258. Q. Dong *et al.*, "15.3 A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2020, pp. 242-244, DOI: 10.1109/ISSCC19947.2020.9062985.
259. C. -X. Xue *et al.*, "16.1 A 22nm 4Mb 8b-Precision ReRAM Computing-in-Memory Macro with 11.91 to 195.7TOPS/W for Tiny AI Edge Devices," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 245-247, DOI: 10.1109/ISSCC42613.2021.9365769.
260. R. Khaddam-Aljameh *et al.*, "HERMES Core – A 14nm CMOS and PCM-based In-Memory Compute Core using an array of 300ps/LSB Linearized CCO-based ADCs and local digital processing," *2021 Symposium on VLSI Technology*, Kyoto, Japan, 2021, pp. 1-2.
261. S. Yin *et al.*, "PIMCA: A 3.4-Mb Programmable In-Memory Computing Accelerator in 28nm for On-Chip DNN Inference," *2021 Symposium on VLSI Circuits*, 2021, pp. 1-2. DOI: 10.23919/VLSICircuits52068.2021.9492403.
262. G. Yuan *et al.*, "FORMS: Fine-grained Polarized ReRAM-based In-situ Computation for Mixed-signal DNN Accelerator," *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, Valencia, Spain, 2021, pp. 265-278, DOI: 10.1109/ISCA52012.2021.00029.
263. H. Caminal *et al.*, "CAPE: A Content-Addressable Processing Engine," *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Seoul, Korea (South), 2021, pp. 557-569, DOI: 10.1109/HPCA51647.2021.00054.

 doi:10.20944/preprints202407.0025.v2

264. S. Lee et al., "A 1ynm 1.25V 8Gb, 16Gb/s/pin GDDR6-based Accelerator-inMemory supporting 1 TFLOPS MAC Operation and Various Activation Functions for Deep-Learning Applications", SK hynix, ISSCC, Feb 2022. DOI: 10.1109/ISSCC42614.2022.9731711.
265. H. Fujiwara *et al.*, "A 5-nm 254-TOPS/W 221-TOPS/mm2 Fully Digital Computing-in-Memory Macro Supporting Wide-Range Dynamic-Voltage-Frequency Scaling and Simultaneous MAC and Write Operations," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731754.
266. J. -S. Park *et al.*, "A Multi-Mode 8K-MAC HW-Utilization-Aware Neural Processing Unit with a Unified Multi-Precision Datapath in 4nm Flagship Mobile SoC," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 246-248, DOI: 10.1109/ISSCC42614.2022.9731639.
267. H. Zhu *et al.*, "COMB-MCM: Computing-on-Memory-Boundary NN Processor with Bipolar Bitwise Sparsity Optimization for Scalable Multi-Chiplet-Module Edge Machine Learning," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731657.
268. D. Niu *et al.*, "184QPS/W 64Mb/mm23D Logic-to-DRAM Hybrid Bonding with Process-Near-Memory Engine for Recommendation System," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731694.
269. Y. -C. Chiu *et al.*, "A 22nm 4Mb STT-MRAM Data-Encrypted Near-Memory Computation Macro with a 192GB/s Read-and-Decryption Bandwidth and 25.1-55.1TOPS/W 8b MAC for AI Operations," *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 178-180, DOI: 10.1109/ISSCC42614.2022.9731621.
270. W. -S. Khwa *et al.*, "11.3 A 40-nm, 2M-Cell, 8b-Precision, Hybrid SLC-MLC PCM Computing-in-Memory Macro with 20.5 - 65.0TOPS/W for Tiny-Al Edge Devices," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, 2022, pp. 1-3. DOI: 10.1109/ISSCC42614.2022.9731670
271. S. D. Spetalnick *et al.*, "A 40nm 64kb 26.56TOPS/W 2.37Mb/mm2RRAM Binary/Compute-in-Memory Macro with 4.23x Improvement in Density and >75% Use of Sensing Dynamic Range," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731725.
272. M. Chang *et al.*, "A 40nm 60.64TOPS/W ECC-Capable Compute-in-Memory/Digital 2.25MB/768KB RRAM/SRAM System with Embedded Cortex M3 Microprocessor for Edge Recommendation Systems," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731679.
273. D. Wang et al., "DIMC: 2219TOPS/W 2569F2/b Digital In-Memory Computing Macro in 28nm Based on Approximate Arithmetic Hardware," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 266-268, DOI: 10.1109/ISSCC42614.2022.9731659.
274. D. Wang et al., "DIMC: 2219TOPS/W 2569F2/b Digital In-Memory Computing Macro in 28nm Based on Approximate Arithmetic Hardware," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 266-268, DOI: 10.1109/ISSCC42614.2022.9731659.
275. J. -M. Hung *et al.*, "An 8-Mb DC-Current-Free Binary-to-8b Precision ReRAM Nonvolatile Computing-in-Memory Macro using Time-Space-Readout with 1286.4-21.6TOPS/W for Edge-AI Devices," *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731715.
276. Yue *et al.*, "15.2 A 2.75-to-75.9TOPS/W Computing-in-Memory NN Processor Supporting Set-Associate Block-Wise Zero Skipping and Ping-Pong CIM with Simultaneous Computation and Weight Updating," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, 2021, pp. 238-240. DOI: 10.1109/ISSCC42613.2021.9365958.

277. J. Yue *et al.*, "14.3 A 65nm Computing-in-Memory-Based CNN Processor with 2.9-to-35.8TOPS/W System Energy Efficiency Using Dynamic-Sparsity Performance-Scaling Architecture and Energy-Efficient Inter/Intra-Macro Data Reuse," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2020, pp. 234-236, DOI: 10.1109/ISSCC19947.2020.9062958.

278. Y. Wang *et al.*, "A 28nm 27.5TOPS/W Approximate-Computing-Based Transformer Processor with Asymptotic Sparsity Speculating and Out-of-Order Computing," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 1-3, DOI: 10.1109/ISSCC42614.2022.9731686.

279. J. -S. Park *et al.*, "A Multi-Mode 8K-MAC HW-Utilization-Aware Neural Processing Unit with a Unified Multi-Precision Datapath in 4nm Flagship Mobile SoC," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2022, pp. 246-248, DOI: 10.1109/ISSCC42614.2022.9731639.

280. K. Matsubara *et al.*, "4.2 A 12nm Autonomous-Driving Processor with 60.4TOPS, 13.8TOPS/W CNN Executed by Task-Separated ASIL D Control," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 56-58, DOI: 10.1109/ISSCC42613.2021.9365745.

281. A. Agrawal *et al.*, "9.1 A 7nm 4-Core AI Chip with 25.6TFLOPS Hybrid FP8 Training, 102.4TOPS INT4 Inference and Workload-Aware Throttling," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 144-146, DOI: 10.1109/ISSCC42613.2021.9365791.

282. H. Mo *et al.*, "9.2 A 28nm 12.1TOPS/W Dual-Mode CNN Processor Using Effective-Weight-Based Convolution and Error-Compensation-Based Prediction," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 146-148, doi: 10.1109/ISSCC42613.2021.9365943.

283. J. -S. Park *et al.*, "9.5 A 6K-MAC Feature-Map-Sparsity-Aware Neural Processing Unit in 5nm Flagship Mobile SoC," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 152-154, DOI: 10.1109/ISSCC42613.2021.9365928.

284. R. Eki *et al.*, "9.6 A 1/2.3inch 12.3Mpixel with On-Chip 4.97TOPS/W CNN Processor Back-Illuminated Stacked CMOS Image Sensor," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 154-156, DOI: 10.1109/ISSCC42613.2021.9365965.

285. C. -H. Lin *et al.*, "7.1 A 3.4-to-13.3TOPS/W 3.6TOPS Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2020, pp. 134-136, DOI: 10.1109/ISSCC19947.2020.9063111.

286. C. -H. Lin *et al.*, "7.1 A 3.4-to-13.3TOPS/W 3.6TOPS Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 134-136, DOI: 10.1109/ISSCC19947.2020.9063111.

287. C. -H. Lin *et al.*, "7.1 A 3.4-to-13.3TOPS/W 3.6TOPS Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2020, pp. 134-136, DOI: 10.1109/ISSCC19947.2020.9063111.

288. W. -H. Huang *et al.*, "A Nonvolatile Al-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM Compute-in-Memory Macro and 51.4-251TOPS/W," *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2023, pp. 15-17, DOI: 10.1109/ISSCC42615.2023.10067610.

289. W. -H. Huang *et al.*, "A Nonvolatile Al-Edge Processor with 4MB SLC-MLC Hybrid-Mode ReRAM Compute-in-Memory Macro and 51.4-251TOPS/W," *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2023, pp. 15-17, doi: 10.1109/ISSCC42615.2023.10067610.

290. T. Tambe *et al.*, "22.9 A 12nm 18.1TFLOPs/W Sparse Transformer Processor with Entropy-Based Early Exit, Mixed-Precision Predication and Fine-Grained Power Management," *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2023, pp. 342-344, doi: 10.1109/ISSCC42615.2023.10067817.

 doi:10.20944/preprints202407.0025.v2

291. T. Tambe *et al*., "22.9 A 12nm 18.1TFLOPs/W Sparse Transformer Processor with Entropy-Based Early Exit, Mixed-Precision Predication and Fine-Grained Power Management," *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2023, pp. 342-344, doi: 10.1109/ISSCC42615.2023. 10067817.
292. Y. -C. Chiu *et al*., "A 22nm 8Mb STT-MRAM Near-Memory-Computing Macro with 8b-Precision and 46.4-160.1TOPS/W for Edge-AI Devices," *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2023, pp. 496-498, DOI: 10.1109/ISSCC42615.2023.10067563.
293. G. Desoli *et al*., "16.7 A 40-310TOPS/W SRAM-Based All-Digital Up to 4b In-Memory Computing Multi-Tiled NN Accelerator in FD-SOI 18nm for Deep-Learning Edge Applications," *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2023, pp. 260-262, DOI: 10.1109/ISSCC42615.2023.10067422.
294. G. Desoli *et al*., "16.7 A 40-310TOPS/W SRAM-Based All-Digital Up to 4b In-Memory Computing Multi-Tiled NN Accelerator in FD-SOI 18nm for Deep-Learning Edge Applications," *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2023, pp. 260-262, DOI: 10.1109/ISSCC42615.2023.10067422.
295. MemComputing, "MEMCPU", Accessed on: Jul 1, 2023. Available: https://www.memcpu.com/.
296. IniLabs, "IniLabs", Accessed on: Jul 1, 2023, Available: https://inilabs.com/.
297. Memryx, Accessed on: Aug 1, 2023, Available: https://memryx.com/products/.
298. A. Tavanaei et al., "Deep learning in spiking neural networks", Neural networks. 111:47-63. Mar 2019. 10.1016/j.neunet.2018.12.002.
299. D. S. Modha, et al., "IBM NorthPole neural inference machine", HotChips conference, 2023, Aug 27-29, California, USA.
300. S. Dhruvanarayan, V. Bittorf, "MLSoC™ – An Overview," HotChips Conference 2023, California, USA, August 2023.
301. SiMa.ai, Accessed on: Sept. 3, 2023. Available: https://sima.ai/.
302. Z. Tan, Y. Wu, Y. Zhang, H. Shi, W. Zhang, K. Ma, "A scaleable multi-chiplet deep learning accelerator with hub-side 2.5D heterogeneous integration", HotChip Conference' 2023. California, USA, August 2023.
303. E. Mahurin, "Qualcomm Hexagon NPU", HotChip Conference'2023, California, USA, August 2023.
304. Dharmendra S. Modha et al., "Neural Inference at the Frontier of Energy, Space, and Time", Computer Science, 382, pp. 329-335, October, 2023. DOI: 10.1126/science.adh1174.
305. Bill Dally, 'Hardware for Deep Learning', NVIDIA Corporation, HotChip Conference, 2023, California, USA, August 2023.
306. J. H. Kim, Y. Ro, J. So, S. Lee, S.-H. Kang, Y. Cho, H. Kim, B. Kim, K. Kim, S. Park, J.-S. Kim, S. Cha, W.-J. Lee, J. Jung, J.-G. Lee, J. Lee, J.H. Song, S. Lee, J. Cho, J. Yu, and K. Sohn, 'Samsung PIM/PNM for Transformer based AI : Energy Efficiency on PIM/PNM Cluster', HotChip Conference'2023, California, USA, August 2023.
307. Ambarella, Accessed on : March 5, 2024, Available: https://www.ambarella.com/products/iot-industrial-robotics/.
308. W-S Khwa, P-C Wu, J-J Wu, J-W Su, H-Y Chen, Z-E Ke, T-C Chiu, J-M Hsu, C-Y Cheng, Y-C Chen, C-C Lo, R-S Liu, C-C Hsieh,
K-T Tang, M-F Chang, 'A 16nm 96Kb Integer/Floating-Point Dual Mode-Gain-CellComputing-in-Memory Macro Achieving 73.3 163.3TOPS/W and 33.2-91.2TFLOPS/W for AI-Edge Devices' *2024 IEEE International Solid- State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2024.

 doi:10.20944/preprints202407.0025.v2

309. M-E Shih, S-W Hsieh, P-Y Tsa, M-H Lin, P-K Tsung, E-J Chang, J Liang, S-H Chang,C-L Huang, Y-Y Nian, Z Wan, S Kumar, C-X Xue, G Jedhe, H. Fujiwara, H Mori, C-Wei Chen, P-H Huang, C-F Juan1, C-Y Chen, T-Y Lin, CH Wang, C-C Chen, K Jou, 'NVE: A 3nm 23.2TOPS/W 12b-Digital-CIM-Based Neural Engine for High Resolution Visual-Quality Enhancement on Smart Devices', *2024 IEEE International Solid- State Circuits Conference (ISSCC),* San Francisco, CA, USA, 2024.

310. K Nose, T Fujii, K Togawa, S Okumura, K Mikami, D Hayashi, T Tanaka, T Toi, 'A 23.9TOPS/W @ 0.8V, 130TOPS AI Acceleratorwith 16× Performanc e-Accelerable Pruning in 14nm Heterogeneous Embedded MPU for Real-Time Robot Applications', *2024 IEEE International Solid- State Circuits Conference (ISSCC),* San Francisco, CA, USA, 2024