algorithms

IMPACT FACTOR 2.1

CITESCORE 4.5

Article

Recent Developments in Low-Power AI Accelerators: A Survey

Christoffer Åleskog, Håkan Grahn and Anton Borg

Topic Collection

Parallel and Distributed Computing: Algorithms and Applications

Edited by

Prof. Dr. Charalampos Konstantopoulos and Prof. Dr. Grammati Pantziou

MDPI

https://doi.org/10.3390/a15110419

algorithms

MDPI

Article

Recent Developments in Low-Power AI Accelerators: A Survey

Christoffer Åleskog *, Håkan Grahn and Anton Borg

Department of Computer Science, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden
* Correspondence: christoffer.aleskog@bth.se

Abstract: As machine learning and AI continue to rapidly develop, and with the ever-closer end of Moore's law, new avenues and novel ideas in architecture design are being created and utilized. One avenue is accelerating AI as close to the user as possible, i.e., at the edge, to reduce latency and increase performance. Therefore, researchers have developed low-power AI accelerators, designed specifically to accelerate machine learning and AI at edge devices. In this paper, we present an overview of low-power AI accelerators between 2019–2022. Low-power AI accelerators are defined in this paper based on their acceleration target and power consumption. In this survey, 79 low-power AI accelerators are presented and discussed. The reviewed accelerators are discussed based on five criteria: (i) power, performance, and power efficiency, (ii) acceleration targets, (iii) arithmetic precision, (iv) neuromorphic accelerators, and (v) industry vs. academic accelerators. CNNs and DNNs are the most popular accelerator targets, while Transformers and SNNs are on the rise.

Keywords: survey; hardware accelerator; low-power; performance; machine learning; artificial intelligence; neural networks

Citation: Åleskog, C.; Grahn, H.; Borg, A. Recent Developments in Low-Power AI Accelerators: A Survey. *Algorithms* **2022**, *15*, 419. https://doi.org/10.3390/a15110419

Academic Editors: Charalampos Konstantopoulos and Grammati Pantziou

Received: 30 September 2022
Accepted: 3 November 2022
Published: 8 November 2022

1. Introduction

Deep learning is a popular machine learning approach, with many application areas [1,2]. However, advances in technology and the ending of Moore's Law [3,4] have resulted in a rise in hardware devices specifically designed to accelerate Artificial Intelligence (AI) and Deep Neural Networks (DNN), with a strong focus on Convolutional Neural Networks (CNN) [5]. Both DNNs and CNNs are prevalent in acceleration due to their numerous uses in a variety of fields (e.g., robotics, object detection, video processing, branch prediction [1]) and their use of matrix operations, which is easy to parallelize.

As accelerator architectures continue to evolve and develop, and even commercialize, the demand for efficient accelerators continues to grow. A focus on accelerating machine learning directly at the edge has emerged, e.g., in low-power devices such as cellphones or cameras, adding concerns for privacy and an ever-increasing demand for higher performance [2]. With stricter power requirements, ranging from 1 mW [6] to $\leq$10 W, new strategies to accommodate the limitation of both power and area are necessary.

This paper presents recent advances in low-power AI accelerators over the last four years. Unlike previous surveys on AI accelerators [2,5,7–9], this survey focuses on a range of different types of accelerators (e.g., CNN, DNN, LSTM, Transformer) and with an upper bound on power of $\leq$10 W. The paper presents and discusses the findings using five different aspects. AI accelerators are investigated in terms of (i) power, performance, and power-efficiency, (ii) the types of models accelerated, i.e., acceleration targets, (iii) precision and number formats supported, (iv) neuromorphic accelerators, and finally (v) accelerators developed by companies.

The paper is organized as follow. Section 2 presents the background and related work. This is followed by Section 3, which contains the aims and scope, and the methodology used. Section 4 presents the findings. This section is divided into several subsections, detailing the different aspects, e.g., power efficiency or accelerators by companies. Finally, the findings are discussed in Section 5, and our conclusions are presented in Section 6.

2. Background and Related Work

The idea of the accelerator is to be able to increase the performance and lower the power consumption by utilizing optimization strategies specific to the problem that the hardware are meant to solve. The most common of these strategies is parallelization; the idea that multiple calculations, independent of each other, can be done in parallel.

The question is then: What applications are parallelizable and are in a dire need of increased performance and decreased power consumption. AI, specifically neural networks in the context of ML, is one answer [2].

Presented in this section is a short overview of the core mechanics for acceleration in AI/ML models; what they are used for and how they can be accelerated. Following that, a brief description of a common structure of an accelerator and how it applies the knowledge learned from the core mechanics to be able to accelerate AI/ML models are presented. Finally, other surveys on AI accelerators are briefly summarized, and the difference between their and our work is presented and explained.

2.1. Core Operation in Artificial Neural Networks

The core operation found in all types of neural networks (CNN, DNN, LSTM, RNN, Transformer, etc.) is the matrix multiplication and accumulation (MAC) operation, excluding Spiking Neural Networks (SNN).

For example, there are two common layers in all CNN models; convolution (CONV) and fully connected (FC). In both layers, the scalar product (MAC) is the core in their computations. Both layers use the scalar product to multiply inputs with weights (filter elements for the convolution layer), and then accumulate the results into one value. In the CONV layer, a sliding window (filter) is moved over the input data, calculating the scalar product between the selected inputs and the filter. Similarly, one FC layer is constructed with multiple neurons, each of which calculates the scalar product between the inputs and the weights, adding a bias and then using the result in the activation function to produce an output. As a demonstration, the calculation of one neuron in the FC layer of an artificial neural network (ANN) is presented in Equation (1) and illustrated in Figure 1.

$$y_j = f\left(\sum_{i=1}^{N} w_{ji} \times x_i + b_j\right), \tag{1}$$

where w is a matrix of weights, x is a vector of inputs, b correspond to the bias, f refer to the activation function, N is the number of inputs to the neuron, and y is the neuron output. Because each neuron is computed with Equation (1) and independent of other neurons in the same layer, it enables parallelization of the computation between neurons in a layer. This is achieved by computing both the neurons and the individual input-weight pairs in parallel (accumulating the results in the process).

The parallelization of the CONV layer is done in a similar way; computing each frame of the sliding window with the corresponding inputs and the individual input–filter weight pairs in parallel (also accumulating the results in the process). A similar process, using the MAC operation, is used for other types of ANNs.

2.2. Typical Accelerator Structure

To understand how the parallelization of the MAC operations is achieved in most accelerators, a generic, typical accelerator architecture is illustrated in Figure 2.

Weights and inputs (that may be inputs to the network itself or activation results from a previous layer) move through the grid of processing elements (PEs) in a rhythmical fashion. In the context of the FC layer in an ANN, a row (or column, depending on implementation) represents one neuron. This grid of simple arithmetic logical units (ALUs, included in each PE) is referred to as a systolic array, and is the core foundation of all accelerators. The choice and placement of buffers (input, weight, and output buffers in Figure 2) varies from accelerator to accelerator. However, the idea is basically the same; input data and

weights are loaded into their respective buffers and propagate through the systolic array, which computes the multiplication and then accumulates the results (either in the PEs themselves or at the end of a lane of PEs, depending on the dataflow used).

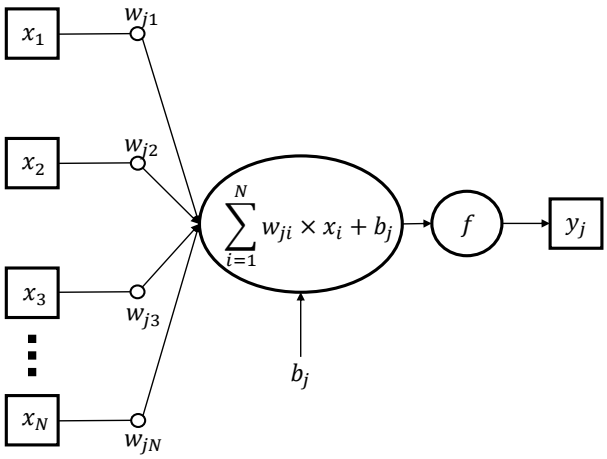

Figure 1. Graph diagram of a single neuron in a traditional ANN.

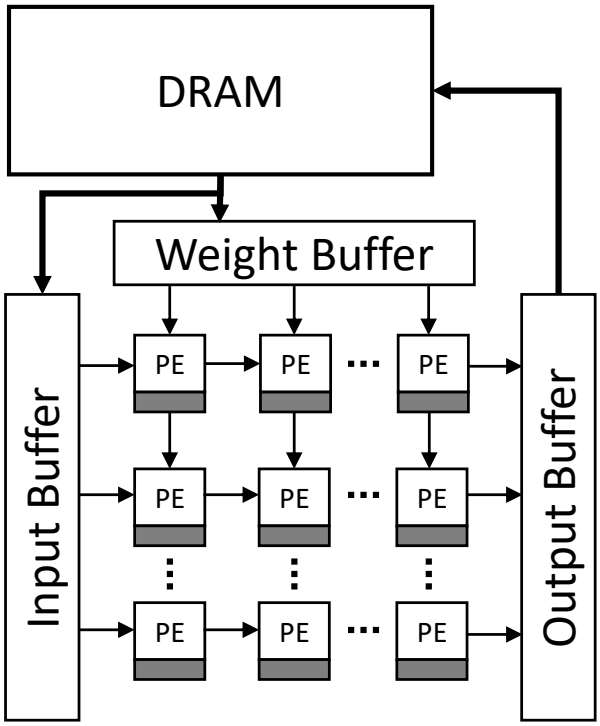

Figure 2. Typical architecture of an ANN accelerator. Each PE has a private register attached to them (gray box), used for storing partial sums. The structure of the grid of PEs varies depending on dataflow used.

Many optimization strategies exist for these types of accelerators, e.g., pruning (reduce number of operations), quantization (reduced precision, either fixed or dynamic), dataflows (Weight Stationary, Output Stationary, No Local Reuse, etc.), Near-Memory Processing (NMP), Compute-In-Memory (CIM), and compact architectures (reduce filter sizes and number of weights). For a more detailed explanation of the mentioned strategies and how DNNs are accelerated in hardware, we recommend the paper by Sze et al. [10]. Further, reducing the overhead of parallelization for DNN AI accelerators can be done via different interconnection strategies. An overview of this topic has been written by Nabavinejad et al. [11].

2.3. Related Work

In recent years, the focus on research for domain-specific hardware accelerators, specifically to accelerate DNNs, has increased drastically. With notable AI accelerators like ICT's DianNao [12] and its successors (DaDianNao [13], ShiDianNao [14], PuDianNao [15]), IBM's TrueNorth [16,17], and MIT's Eyeriss [18] paving the way for the future of AI accelerator research. Further, approaches for validating and testing implemented AI accelerators is an important aspect, see, e.g., [19] which proposes an test architecture for low-power ASIC AI accelerators. However, this is considered out-of-scope for this survey.

Different surveys discussing and analyzing the growing archive of AI accelerators, both on the research and the commercial side, have been published over the years. The following paragraphs go through some of these published surveys.

Reuther et al. [5] surveyed DNN and CNN accelerators published or announced between 2014 and 2019 (both commercial and for research purposes). They analyze the accelerators performance, numerical precision, power consumption, and inference versus training. The surveyed accelerators include CPU-based and GPU-based processors, and ASIC, FPGA, and dataflow accelerators. Reuther et al. observed that many of the previous processor trends are no longer increasing, resulting in the increased popularity of domain-specific hardware accelerators. They observed that recent accelerators had a power utilization between 10–300 W, and that processors/accelerators below 100 W focused solely on inference. They noted that most accelerators are able to reduce their numerical precision by half (16-bit) or even use single bit representations, thereby increasing operations per second without impacting the accuracy overly much. Reuther et al. updated the survey in the following two years [7,8]; adding more accelerators (the former focusing on publicly announced accelerators and the latter focusing solely on commercial accelerators) and including new architectures like neuromorphic designs [20,21], flash-based analog memory processing [22], and photonic-based processing [23].

Talib et al. [24] published a systematic literature review on hardware implementation for different AI algorithms. Reviewing 169 different research papers, published between the years 2009 and 2019, focusing on ASIC, FPGA, and GPU based hardware accelerator implementations. They observed that the majority were FPGA-based implementations, focusing mostly on accelerating CNNs for object detection, with the GPU-based in second place (third was the number of FPGA-based solutions). Talib et al. studied what tools, platforms, and approaches were used in the last decade, and how the performance changes depending on clock frequency and numerical precision.

Li et al. [25] published a short systematic literature review on ANN accelerators with a focus on DNN/CNN co-processors; CPUs integrated with the co-processor acting as the acceleration module for ANN acceleration. They observed that the representative numerical precision formats used were 8-bit integer and 16-bit floating point, and that the throughput varies from 100 GOPS to 10 TOPS.

Lin et al. [9] lists and reviews CNN accelerators for image recognition, focusing on ASIC, FPGA, and coarse-grained reconfigurable array (CGRA) implementations. They observe that CGRA uses 16-bit to 32-bit fixed point precision formats, while 16-bit floating point or fixed point is preferred for ASIC and FPGAs.

Seo et al. [22] surveyed all-digital and analog/mixed-signal AI accelerators. They discuss and present the common optimization strategies used in digital AI accelerators. Continuing with a deep look into analog/mixed-signal accelerators; discussing what they are and how they work. Seo et al. conclude the paper by presenting the emerging trends for both all-digital and analog/mixed-signal AI accelerators. Digital trends include reconfigurable MAC arrays, dynamic mixed precision support, sparsity awareness for weights and activations (pruning), and the support of a variety of AI models. On the other hand, analog/mixed-signal trends include many-macro CIM, reduced overhead of ADC, increased DNN accuracy, and sparsity awareness.

This survey differs from, and extends, previous surveys in three ways. First, it focuses solely on low-power ($\leq$10 W) AI accelerators published in the last four years (2019–2022).

Second, previous surveys have mainly focused on a specific domain or a specific accelerator target, e.g., CNN. This survey is not limited to AI accelerators based on ANNs, nor a specific domain, providing a broader overview of low-power AI accelerator research. Third, it investigates the relationship between throughput, power, and power efficiency. Something that has not been explicitly investigated in the previous research.

3. Method

3.1. Aim and Scope

We used a backward snowballing approach [26] with the initial papers taken from two major conferences (ISCA and ASPLOS) in computer architecture and one workshop (AccML) dedicated to ML acceleration. Under the snowballing processes, papers outside the scope of the initial set were considered, including company affiliated accelerators (data gathered from the companies own websites). However, papers outside the snowballing process were not the focus under the searching process, hence the limited number of company affiliated accelerators in our survey. The papers (including company briefs, white papers, and other information gathered from companies' websites) were accepted based on certain criteria defined in Section 3.2. Therefore, the scope of our survey is on the current and emerging research in domain-specific hardware accelerators. As with any systematic literature method, there is no guarantee that all papers in the domain were accounted for. However, we think that our chosen method, based on our initial set, sufficiently represents the current research in the field to an acceptable degree for this survey.

3.2. Criterias for Acceptance

The criterias for accepting an article were based on (i) the year it was published, (ii) if it introduced an AI accelerator, and (iii) if it could be counted as low-power. The year it was published is self explanatory; we only accepted papers published between the years 2019 and 2022.

Regarding if an accelerator accelerated AI or not, we based our decisions on our own experiences and knowledge of current and emerging research in AI. In this survey, the AI applications/algorithms are referred to as the acceleration targets. The total number of targets found from the surveyed accelerators are presented in Table 1. Accelerators that accelerate General Matrix Multiplications (GEMM) have been included due to the prevalent use of matrix multiplication in a variety of AI algorithms, including DNNs. Consequentially, MAC and other types of matrix algorithms were included for similar reasons. We count Graph Mining and Personalized Recommendation (PR) systems as AI applications too. As such, the acceleration targets are as follows: ANN, Multilayer Perceptron (MLP), DNN, CNN, Generative Adversarial Network (GAN), GEMM (divided into sparse and dense), PR system, Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Transformer, MAC, Sparse Matrix Transposition, Matrix Algorithms, Graph Mining, and SNN.

We set the upper bound on power to be less than 10 W. The maximum power of a low-power device ($\leq$10 W) in this paper was derived from reviewing a variety of research papers on AI accelerators. A higher value of power was not chosen to take into consideration the power that other components of a device might need, i.e., camera, extra memory, and sensors.

Low-power AI accelerators where the authors are affiliated with a company are also considered. The same criteria as mentioned before were used, except for the mentioned conferences and workshop, company affiliated accelerators from known AI accelerator vendors were also included in this survey. We searched for vendors that we know design accelerators specific for AI acceleration.

Table 1. Number of acceleration targets per year. Divided into company and non-company (research) acceleration targets. Note that an accelerator can accelerate more than one target.

Targets	**2019**		**2020**		**2021**		**2022**	
	Research	**Company**	**Research**	**Company**	**Research**	**Company**	**Research**	**Company**
CNN	10	4	5	4	7	6	2	-
GAN	1	-	1	-	1	-	1	-
DNN	2	3	4	-	2	5	1	-
Transformer	1	-	1	-	3	2	3	-
SpGEMM	1	-	2	-	1	-	-	-
SNN	1	-	3	-	2	-	2	-
ANN	1	-	1	-	-	-	-	-
RNN	2	-	-	4	1	3	1	-
LSTM	1	1	-	4	-	3	-	-
PR system	-	-	1	-	-	-	-	-
GEMM	-	-	2	-	3	-	1	-
MAC	-	-	1	-	-	-	-	-
Matrix Alg.	-	-	1	-	-	-	-	-
MLP	-	-	-	-	1	-	-	-
Graph Mining	-	-	-	-	-	-	2	-
SpMatrix Transposition	-	-	-	-	-	-	1	-
Attention	-	-	-	-	-	-	1	-
Unique targets	9	3	11	3	9	5	10	0
Total targets	19	8	22	12	21	19	15	0

As presented in Table 2, there are more accelerators presented in academic research papers than accelerators where the authors are affiliated with a company. This will be reflected in the result, mostly when analyzing the company accelerators.

Table 2. Number of accelerators per year. Divided into company and non-company (research) accelerators.

	2019	**2020**	**2021**	**2022**
Company	6	4	8	0
Academia	16	17	16	12
Total	22	21	24	12

3.3. Missing Data

Since there exist no standards on what information should exist in both research papers and company briefs for AI accelerators, a few metrics and measurements are missing from the gathered information regarding different AI accelerators in our survey, e.g., power, throughput, power efficiency, etc.

There are times when this could be rectified. In the data gathered from the surveyed accelerators, we are able to compute some of these missing values. Throughput, or operations per second, can be calculated with Equation (2).

$$Throughput = \#MACs \times Frequency = \#MACs \times \frac{Cycles}{Second}, \tag{2}$$

where $\#MACs$ is the number of MAC components in the accelerator (often calculated from the number of PEs used and their internal components), *Frequency* is the frequency in Hz, and *Throughput* is the number of operations per second, often translated to Giga Operations per Second (GOPS) or Tera Operations per Second (TOPS). A similar approach is used for calculating the throughput used in SNN accelerators, where Giga-synaptic Operations per Second (GSyOPS) is often preferred.

If power is missing, we can calculate it by using Equation (3).

$$Power = \frac{Energy}{Cycle} \times Frequency = \frac{Energy}{Cycle} \times \frac{Cycles}{Second}, \quad (3)$$

where the *Energy* of each component of the accelerator (if mentioned) is measured in J per cycle (operation). The frequency is measured in Hz, i.e., $cycles/s$, and the result is *Power*, measured in J/s = W.

We can calculate power efficiency (GOPS/W) if both throughput and power are mentioned or if they can be calculated. This is done by dividing the throughput with the power. Area efficiency is calculated in a similar fashion.

4. Results

This section is divided into eight parts, each covering different aspects of low-power AI accelerators. Section 4.1 presents the results regarding power, throughput, and power efficiency, while power vs. area is addressed in Section 4.2. Section 4.3 discusses common approaches to reduce power in AI accelerators. In Section 4.4, we present which type of AI models the accelerators target. In Section 4.5, we provide an overview of the different number formats and precision the accelerators support. Section 4.6 presents neuromorphic accelerators, and Section 4.7 presents a focused view of accelerators developed by companies. Finally, in Section 4.8, we summarize our findings.

As a foundation to base our results on, we compare our data with the conclusions and findings found in the surveys by Reuther et al. [5,7,8]. In Appendix A, and in Tables A1–A4, all reviewed low-power AI accelerators from the last four years are presented.

4.1. Power, Throughput, and Power Efficiency

We start our survey of low-power AI accelerators with a comparison of the power consumption and the throughput. In Figure 3, we have plotted the power consumption (measured in mW) vs. the number of operations performed per second (counted as Giga Operations Per Second, GOPS). We have included three lines in the figure that correspond to a power efficiency of 10 TOPS/W, 1 TOPS/W, and 100 GOPS/W, respectively. The different colors in the figure correspond to the implementation technology, i.e., ASIC (red), FPGA (green), or if the particular design is only simulated (blue). In addition, an empty circle refers to an accelerator that only supports inference, while a filled circle refers to an accelerator that also supports training.

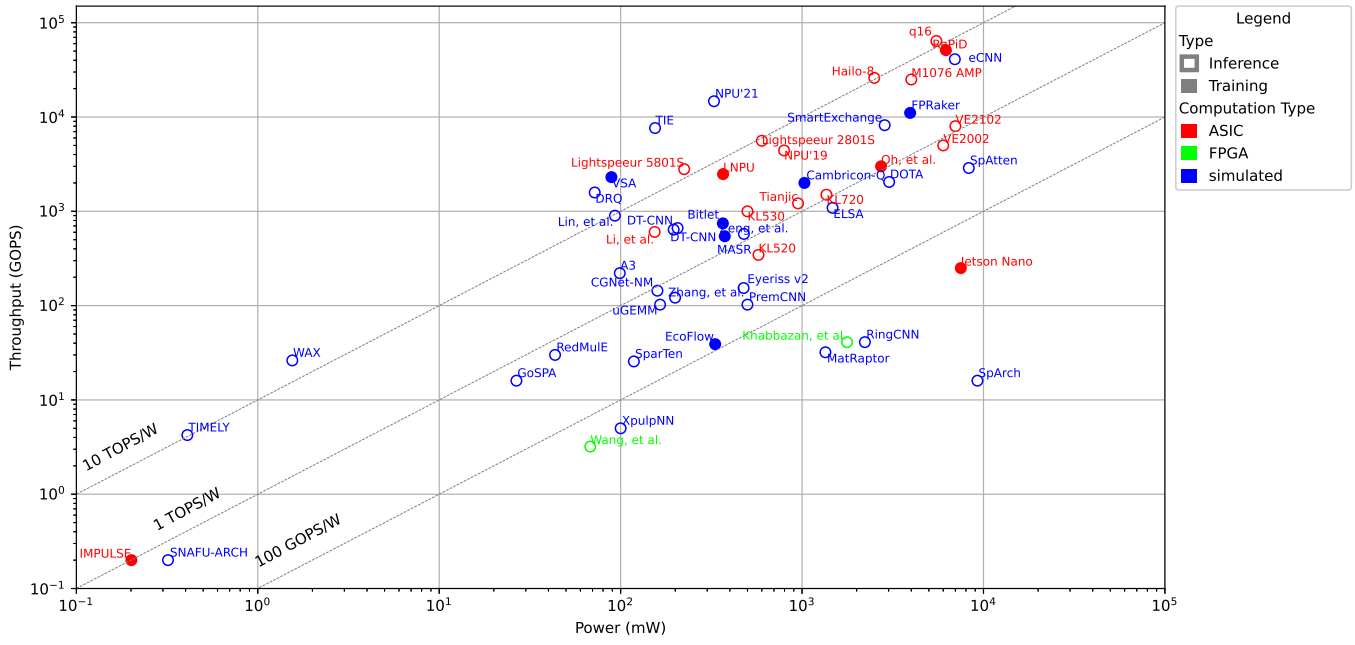

Figure 3. Power (mW) vs. throughput (GOPS) scatter plot of low-power AI accelerators at logarithmic scales.

Compared to a similar graph by Reuther et al. [5,7,8], we can observe that the recent accelerators stay mostly around the trendlines for a power efficiency between 100 GOPS/W and 10 TOPS/W, which is inline with previous studies. There is, however, a homogeneity we can see in the data. The low-power accelerators tend to have a power consumption above ≈100 mW and a throughput above ≈1 GOPS; few accelerators lay beneath these thresholds. Further, we can observe that ASIC implementations in general have a better power efficiency than FPGA implementations, which is expected. Finally, we can observe that the majority of the surveyed accelerators target only inference. A possible explanation for that is that many low-power accelerators target edge or mobile devices, where a pre-trained model is used and only inference needs to be done.

A question to ask is whether the type of target for acceleration is responsible for these deviations in the data compared. For example, SNAFU-ARCH [6] and RedMulE [27] accelerate GEMMs, and IMPULSE [28] accelerates SNNs, while others like WAX [29], GoSPA [30], Wang et al. [31], and TIMELY [32], accelerate CNNs or DNNs. Looking into more detail of SNAFU-ARCH [6], an Ultra-Low-Power (ULP) accelerator and framework for GEMM, we find that it saves energy by configuring PEs and routers for a single operation to minimize switching activity (e.g., specialized PE for basic ALU, multiplier, and memory), resulting in very low power (below 1 mW) and an extremely small area footprint (less than 1 mm^2). From the data in Figure 3, we cannot observe any general trends based on the acceleration target. Hence, we conclude that the type of acceleration target has no significant, general effect on the throughput and power.

In Figure 4, we show for each accelerator which year it was published. The colors in Figure 4 represent the publication years, i.e., 2019 (red), 2020 (green), 2021 (blue), and 2022 (purple). We observe that the throughput and power of low-power AI accelerators have not changed significantly or systematically over the years. Almost all accelerators are placed between the 10 TOPS/W and the 100 GOPS/W lines. Further, we cannot see any clear trend that more recent accelerator designs are more power efficient than earlier designs, i.e., they are not generally closer to the 10 TOPS/W line. A conclusion to draw from that data is that no drastically new innovations have been made that significantly affects the throughput, power, or power efficiency as compared to previous years.

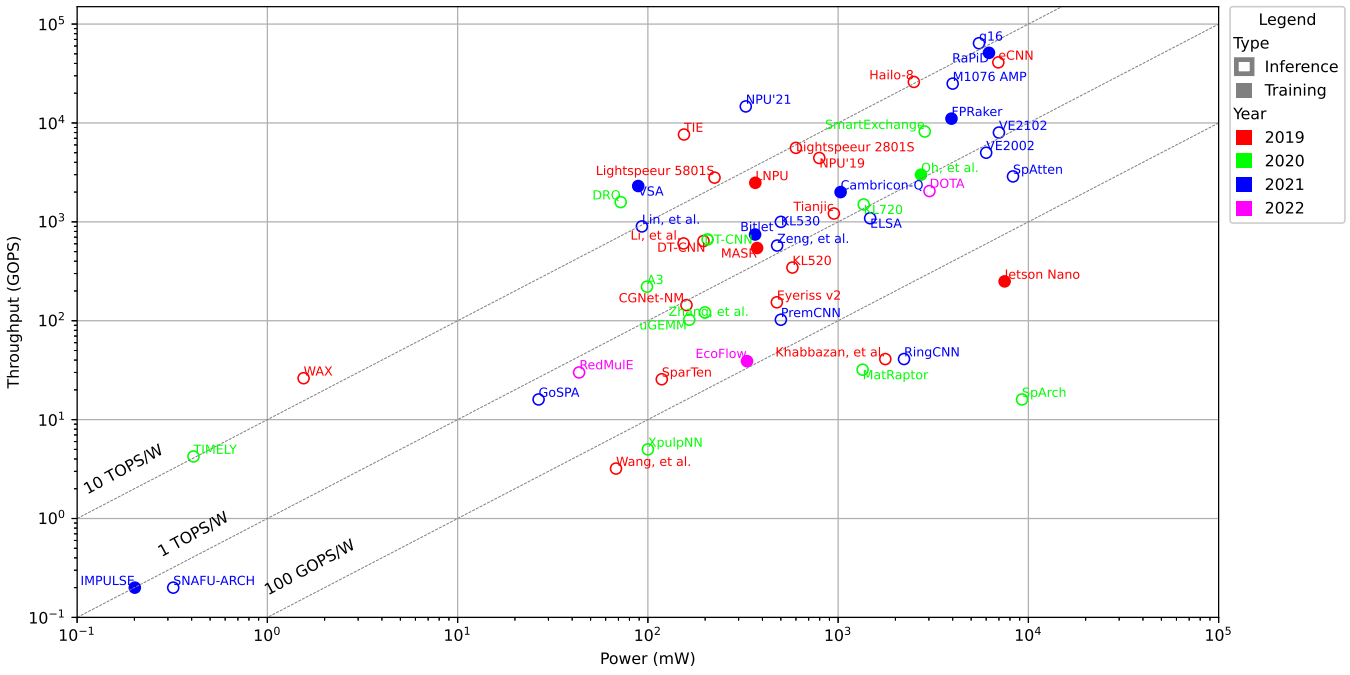

Figure 4. Power (mW) vs. throughput (GOPS) scatter plot of low-power AI accelerators at logarithmic scales. The colors represent the year and the shapes (filled or hollow) represent if it can be used for training and/or inference.

We had an initial hypothesis that low-power AI accelerators designed for training would use more energy than accelerators that are exclusively designed for inference. However, the data presented in Figure 3 do not support this hypothesis. Looking closer at this deviation from our hypothesis, we can observe that accelerators affiliated with a company mostly follow our hypothesis, as shown in Figure 5 where the colors represent company (green) vs. non-company (red) accelerators, respectively. Accelerators designed for ML training tend to have higher power requirements than most other company affiliated accelerators designed only for inference. However, non-company affiliated accelerators do not follow our hypothesis. Neither does it change over time, i.e., the year it was published does not affect if the accelerator itself would be designed for inference or training. This holds true for both company and non-company affiliated low-power AI accelerators. Further, we have observed that accelerators from companies tend to use more power in exchange for higher throughput, as all accelerators from companies have a throughput above 100 GOPS and a power consumption above 100 mW.

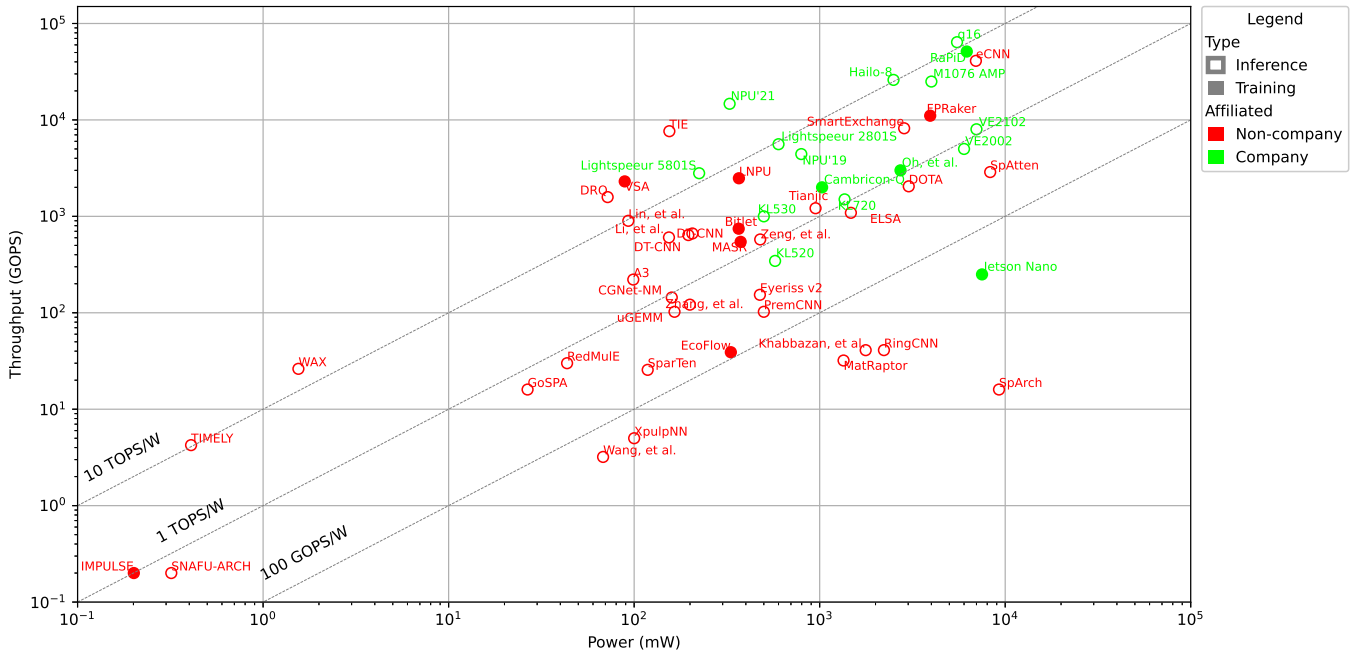

Figure 5. Power (mW) vs. throughput (GOPS) scatter plot of low-power AI accelerators at logarithmic scales. The colors represent an accelerator's affiliation: company (green) or non-company (red) affiliated accelerator.

This observation holds true for power efficiency as well, as shown in the box plot in Figure 6. In Figure 6, the red line denotes the median, the box represents the interval between the first quartile (Q_1) and the third quartile (Q_3). The boundaries of the whiskers are based on 1.5× IQR, where the interquartile range (IQR) is defined as IQR $= Q_3 - Q_1$. The rings outside the whiskers are outliers.

Accelerators from companies have slightly better performance and also tend to be more power efficient. We believe that this observation is true, even though the number of metrics in our data from company affiliated accelerators are not complete. Companies often do not publicly announce all specifications and metrics of their accelerators, which partly explains the low number of data entries from company affiliated accelerators in 2020 (left graph in Figure 6). The mean and median power consumption, performance, and power efficiency for selected groups are presented in Table 3. The selected groups will be presented further in Sections 4.6 and 4.7.

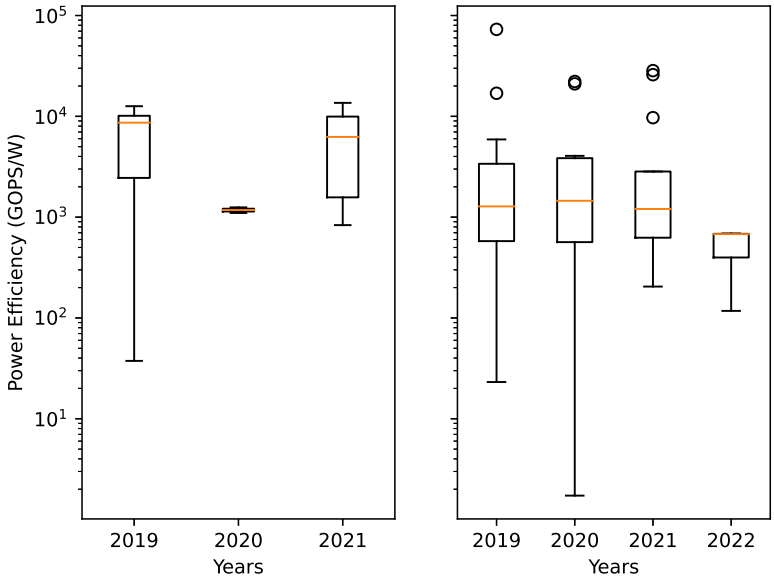

Figure 6. Power efficiency of AI accelerators throughout the years. Left figure: Data from accelerators created by different companies. Right figure: Data from accelerators published as research papers and not affiliated with a company.

Table 3. Power consumption, performance, and power efficiency for selected accelerator groups.

	Power [a]		**Performance** [b]		**Power Efficiency** [c]	
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**
All	1654.6	478.0	5008.6	639.7	6060.7	1264.0
Company	2928.0	1931.2	11,912.1	4423.5	5558.8	4125.0
Non Company	1270.2	206.0	2027.5	148.8	6272.0	1020.7
SNN	938.9	155.8	1172.8	1214.1	9389.3	1278.0
Non SNN	1748.5	500.0	5200.4	622.2	5864.9	1250.0

[a] Power is measured in mW. [b] Performance is measured in GOPS.[c] Power Efficiency is measured in GOPS/W.

4.2. Power, Area, and Clock Frequency

The next aspect that we would like to observe is to take a deeper look into the relationship between power and area. In Figure 7, the ratio of power vs. area over the years is plotted. In Figures 7–9, the red line denotes the median, the box represents the interval between the first quartile (Q_1) and the third quartile (Q_3). The boundaries of the whiskers are based on 1.5× IQR, where the interquartile range (IQR) is defined as IQR $= Q_3 - Q_1$. The rings outside the whiskers are outliers.

In Figure 7, we can observe that there has not been much change with regard to power per square milliliter; there is an overall increase. The reason for this increase is related to the power consumption, i.e., there is an increase in power consumption every year for all accelerators, independent of the accelerator's affiliation.

For the company affiliated accelerators, the increase in power consumption is substantial, as shown in Figure 8, with a peak in 2021 (company affiliated accelerators from 2022 is missing from our data set). It increases every year, excluding NVIDIA's Jetson Nano [33] from 2019, which has the highest overall power consumption reported. Meanwhile, non-company affiliated accelerators have not had such a clear increase in power over the years.

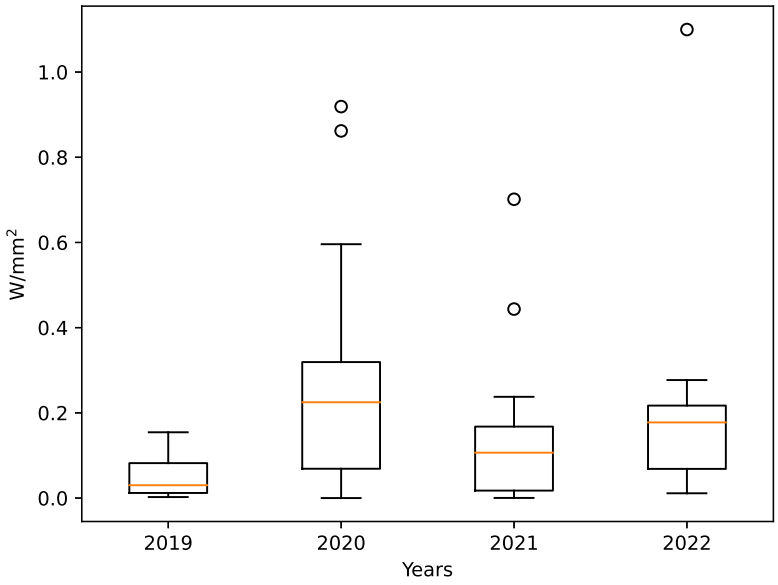

Figure 7. Power per square millimeter over the years.

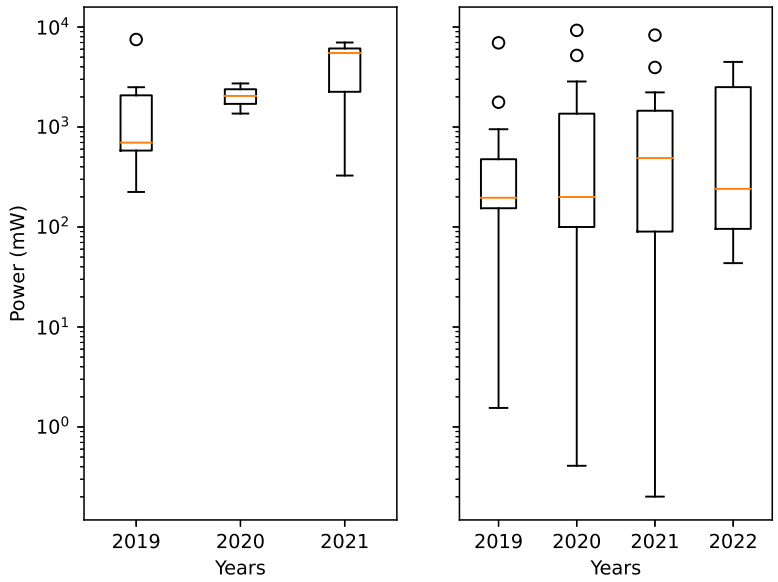

Figure 8. Power consumption throughout the years. Left figure: The data are from accelerators created by different companies. Right figure: The data are from accelerators published as research papers and not affiliated with a company.

Plotting the clock frequencies with regard to year, the cause for the increase in power is uncovered. Depicted in Figure 9, the clock frequencies for both company and non-company affiliated accelerators are plotted. Similar to power, the clock frequencies used in low-power accelerators also increase over time. As the power consumption is proportional to the clock frequency, it follows that higher clock frequencies require more power to operate. This is especially true since Dennard scaling [34] does not hold anymore, thus we cannot obtain higher performance and clock frequencies as we reduce the technology nodes. Therefore, the clock frequency also contributes to the overall increase in power per square millimeter over time.

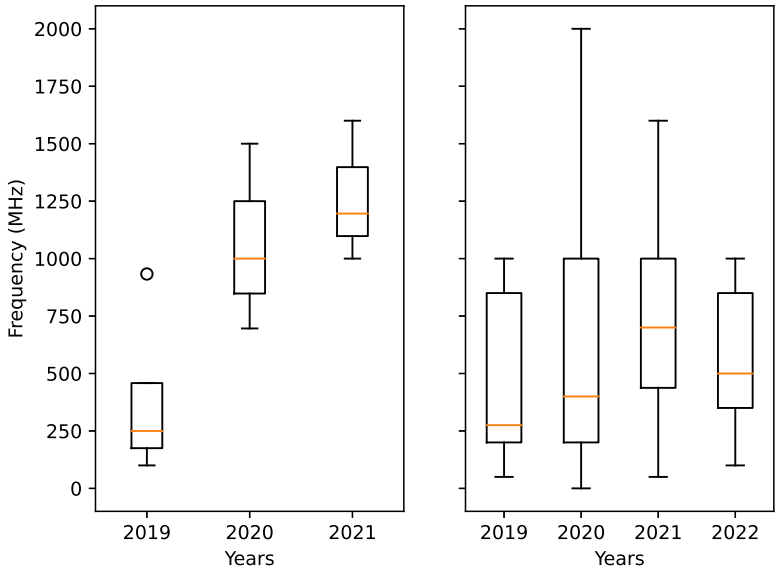

Figure 9. Clock frequencies of AI accelerators throughout the years. Left figure: The data are from accelerators created by different companies. Right figure: The data are from accelerators published as research papers and not affiliated with a company.

Regarding clock frequencies, an observation can be done. The four accelerators that differ significantly from the rest in terms of power (<10 mW), i.e., WAX [29], SNAFU-ARCH [6], TIMELY [32], and IMPULSE [28], also have very low clock frequencies; between 40 MHz and 200 MHz. This shows that low-frequency accelerators use significantly less power at the cost of lower throughput. However, there exist accelerators in the middle of the cluster, such as Zhang et al. [35] and DOTA [36] with clock frequencies of 25 MHz and 100 MHz, that contradict that observation (if only with regards to frequencies). Taking the targets into account, the low-power accelerators that have clock frequencies below or equal to 100 MHz generally accelerate non-ANN targets, e.g., Zhang et al. [35] and DOTA [36] accelerate MACs using CIM and Transformers, respectively. Hence, we can reason that low-power accelerators, if they accelerate general ANNs, mostly have clock frequencies higher than ≈200 MHz to be able to have a power and throughput above ≈200 mW and ≈10 GOPS, respectively. Of course, it is not absolute; other factors, both internal and external, could affect this relationship between power, throughput and frequency.

Area, on the other hand, stays generally the same over the years. The non-company affiliated accelerators' area fluctuates throughout the years, nonetheless staying between 1–100 mm^2. Company affiliated accelerators do report a higher overall area (except in 2020, where they were similar to the non-company affiliated accelerators). Although, this difference in area between company and non-company affiliated accelerators is likely due to the inclusion of other components rather than the core components of the reported accelerator.

We observe, by analyzing the power per square millimeter over the years separately for company and non-company affiliated accelerators, that the company affiliated accelerators do have a lower power-area ratio than the non-company affiliated accelerators. This is probably due to the smaller technology nodes used in their manufacturing (5–28 nm compared with simulated accelerators using 22–65 nm). The dynamic power consumption is proportional to the effective capacitance load, which in a very general reasoning is similar for similar chip areas. Thus, a similar amount of power is needed while increasing the number of transistors per area, thereby increasing throughput. This reasoning matches what is shown in Figure 5, i.e., the company affiliated accelerators have a more consistent throughput vs. power, with the majority having the highest throughput vs. power among all the accelerators.

4.3. Strategies to Decrease Power

Low-power AI accelerators are defined in this survey as AI accelerators with a power consumption of $\leq$10 W. However, how is the decrease in power consumption for low-power AI accelerators achieved compared to other accelerators? In Section 2.2, optimization strategies used in AI accelerators were briefly mentioned, and some of these strategies are used abundantly in low-power AI accelerators to decrease the power consumption as otherwise required. Presented in Table 4 are common strategies used for decreasing power consumption in AI accelerators.

Table 4. Common strategies used to decrease power consumption in AI accelerators.

Strategy	Description	Examples
Pruning	Remove parameters based on different criteria, thereby reducing the workload (e.g., zero valued weights in DNNs and tokens in Transformers can be removed, resulting in less computation).	[37–41]
Quantization	Mapping continuous values to discrete values (i.e., 32-bit float to 16-bit integer), and doing the computations on the discrete representation.	[37,38,42–46]
Dataflow	Reduce movement of data by specifying the flow of data in the architecture (e.g., Weight Stationary, Output Stationary, No Local Reuse, etc.).	[29,47–49]
Near-Memory Processing	Reducing the distance between memory and processing components.	[40,50]
Compute-in-Memory	MAC computation in the analog domain, computed in the memory array. Thereby removing the need for movements of weights.	[28,32]
Reduced flexibility	Reduce the flexibility/reconfigurability of the accelerator by restricting the hardware to perform specific, predetermined tasks (e.g, a PE might only compute a single operation, only support ReLU, etc.).	[6,51]

The descriptions of the different strategies mentioned in Table 4 are quite broad, as there exist many different implementation choices for each one. An example for this is Quantization, where one DNN accelerator might restrict the precision format to 8-bit integers only, while another accelerator might use a more dynamic approach (i.e., dynamically changing the precision format depending on the task at hand). Due to the many different implementations, it is hard to determine which strategy is best. As different implementations and designs can induce different effects on the power consumption. However, it should be noted that all strategies do decrease the power consumption regardless of the implementation used.

4.4. Acceleration Targets

In Table 1, we present all targets that the low-power AI accelerators in this survey accelerate. In total, there are 17 different acceleration targets, divided into nine types of neural networks (CNN, GAN, DNN, ANN, RNN, LSTM, MLP, Transformer, SNN), four matrix algorithms (GEMM, SpGEMM, SpMatrix Transposition, Matrix Alg.), two core mechanisms in neural networks (MAC and Attention), and two graph-based systems (Personalized Recommendation and Graph Mining).

In Figure 10, we show the number of different acceleration targets over the years. We group the targets together, based on similarities in their structures, as follows: RNN refers to RNN and LSTM; DNN includes ANN and MLP; Transformer is grouped with the Attention mechanism; GEMM with SpGEMM; and finally the remaining targets (MAC, SpMatrix Transposition, PR, Graph Mining, Matrix Alg.) are grouped under the label 'Other'.

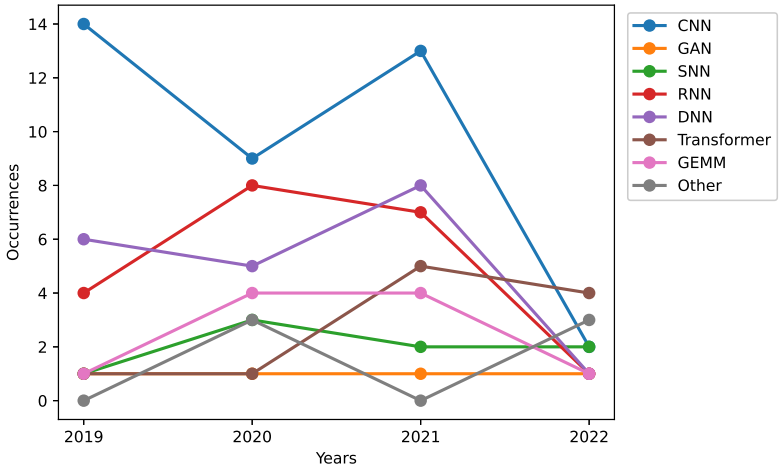

Figure 10. Accelerating targets throughout the years, where RNN refers to RNN and LSTM; DNN includes DNN, ANN, and MLP; Transformer groups the Transformer model and Attention mechanism together; and finally, Other refers to the remaining acceleration targets. Note, however, that some accelerators accelerate a variety of applications; having more than one acceleration target.

We observe that the most common acceleration target is CNN. This has been the case for previous years too, where CNN is the most common choice for accelerating. This stems from the start of the current popularity of neural networks and machine learning in general, as it started with the CNN in the early 2000s [52]. As CNNs are very good at image recognition tasks, and low-power AI accelerators include everything from accelerators in smartphones to complementary components in sensors and cameras, the origin of the CNN's popularity for low-power AI accelerators becomes clear. The second most commonly accelerated targets are DNNs and RNNs, the former can be attributed to the current popularity concerning neural networks in general while the latter to their frequent usage in speech recognition and natural language translation.

A more recent addition to these targets are Transformers, that were introduced in 2017 with the paper by Vaswani et al. [53]. As shown in Figure 10, the use of Transformer as the acceleration target for low-power AI accelerators increased in 2021 compared to previous years. In our opinion, this is attributed to the slow transition into the low-power domain for Transformer models.

An example of another acceleration target is RecNMP [54], a NMP accelerator for accelerating PR (Personalized Recommendation) systems. RecNMP maximally exploits rank-level parallelism and temporal locality of production embedding traces to reduce energy consumption by 45.8% and increase throughput by 4.2 times.

Accelerators can accelerate more than one target application. In the gathered data, 30% of all low-power accelerators used in this survey accelerate more than one target (12 of 79 are from company affiliated accelerators). However, this is not depicted in Figure 10.

Tallying the number of accelerators released per year and comparing it with the number of targets per year, the trend of how common it is to release accelerators that target multiple applications becomes apparent. As observed in Figure 11, the number of targets increases faster than the number of accelerators for company affiliated accelerators. This indicates an increased popularity for accelerators with multiple targets, assuming the trend continues in 2022. A conclusion from this data is that for low-power AI accelerators, domain-specific accelerators become more general over time. However, note that domain-specific accelerators that accelerate multiple targets, often accelerate applications that are similar to each other, e.g., ETHOS-U55 [55] accelerates CNNs, LSTMs, and RNNs. Meaning they accelerate applications that theoretically belong to the same general group of applications, sharing many internal components among them.

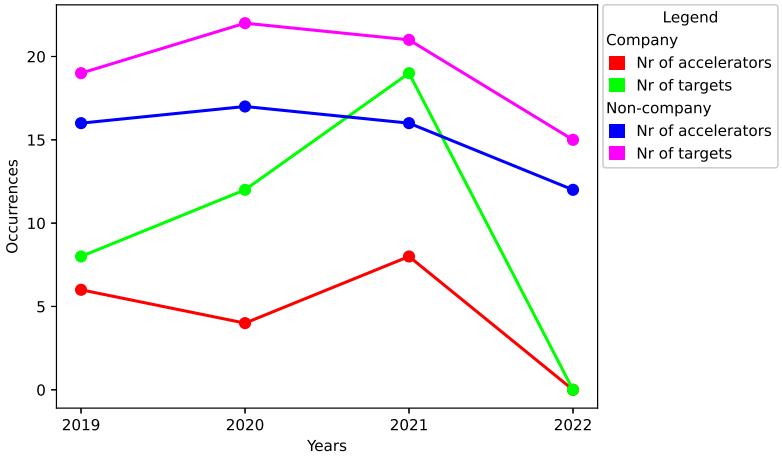

Figure 11. Number of accelerators and targets for both company and non-company affiliated accelerators over the years.

4.5. Number and Precision Formats

Similarly to the acceleration targets, the number and precision formats used for each accelerator are plotted and presented in Figure 12. For similar reasons as for the acceleration targets, the precision formats are grouped together for clarity. The five categories that the accelerators were grouped into are as follows: *INT* denotes the usage of integers, *FP* denotes the usage of floating point formats, *FxP* denotes fixed-point formats, *BFP* refers to Google Brain's bfloat format, *Bit* denotes all accelerators where the number of bits used were mentioned but not the data type (integer, float, fixed-point, etc.), and *unspecified* denotes accelerators where the precision format was not explicitly mentioned, nor the type of the format. It should be noted, however, that similar to acceleration targets, accelerators can use more than one precision format.

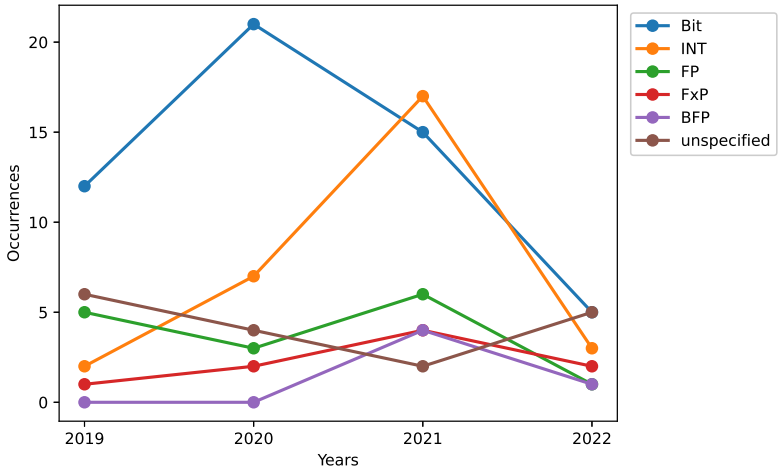

Figure 12. Occurrences of different precision formats throughout the years. Where *unspecified* denotes all accelerators where the precision format used was not explicitly mentioned, and *Bit* denotes all accelerators where the number of bits used were mentioned, but not the type (integer, float, fixed-point, etc.).

Based on the highest number of accelerators that uses a specific format, it is apparent that *Bit* is the most common one. As *Bit* denotes the accelerators that did not specify the type of precision format, we can assume that most, if not all, of these accelerators use integers. This conclusion is based on the fact that most AI accelerators accelerate some kind of neural networks, often models that are quite large. Therefore, integers are often used when accelerating these kind of models, both because it is faster and due to it being

more energy efficient. This hypothesis is backed up by analyzing the results in the survey paper from 2021 by Reuther et al. [8], where they observe that INT8 is the most common precision format, with INT16 and FP16 as close runners-up. Inspecting only accelerators with a power of less than 10 W, it becomes apparent that INT8 and INT16 are the only precision formats used in the multiplication and accumulation for low-power accelerators gathered in surveys by Reuther et al. [8]. Following this observation and assuming that our assumption of the *Bit* format is correct, the integers dominate the field.

The second most common format, ignoring the *unspecified* formats, is the floating point format (*FP*), as seen in Figure 12. According to our results, the use of floating points has not changed much over the years. On the other hand, for *FxP* and *BFP*, a clear increase in usage over time is observed. Regarding the *FxP* format, there is a steady increase of its usage in low-power accelerators. An underlying reason is probably the stricter power budget in low-power AI accelerators. Looking at the *BFP* format, an observation can be made. The popularity of the bfloat increases significantly in 2021, and as bfloat was first announced in 2018, we can assume that the underlying reason for this increase now, instead of before, is due to the transition time into the low-power domain.

Presented in Figure 13, the precision formats are grouped together with regard to the number of bits used. Divided into less or equal to 8-bit precision and above 8-bit precision, it is observed that most accelerators use less than or equal to 8-bit precision formats, which is in line with the results from the survey by Reuther et al. [8]. One might reason that the trend is that more accelerators tend to use $\leq$8-bit precision formats over $>$8-bit precision formats over time, e.g., as a result of power constraints. However, we have observed that $\leq$8-bit precision formats are often complemented by $>$8-bit precision formats, or at least 16-bit precision formats.

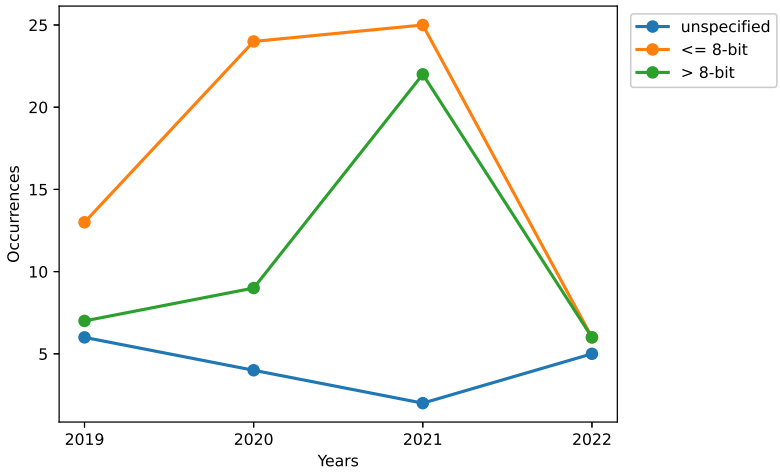

Figure 13. Occurrences of different precision formats throughout the years. Where *$\leq$8-bit* and *$>$8-bit* denote precision formats for less then or equal to 8-bits and greater then 8-bits, respectively. *Unspecified* denotes all accelerators where the precision format used was not explicitly mentioned.

Looking at Figure 13, there is a large increase in higher precision formats in 2021 (more bits). Analyzing the company affiliated accelerators for this year, we see that more than half of them use mixed precisions; 11 use a mix of high and low number of bits in their precision formats, two use only a high number of bits (FP16 for Jetson Nano [33] and Oh et al. [56]), two use only a low number of bits (8-bit for KL520 [57] and KL720 [58]), and the rest (Lightspeeur 5801S [59] and Lightspeeur 2801S [60]) did not specify which precision format their accelerators support. Due to the latter, the number of company affiliated accelerators in 2021 is the sum of the two previous years, and this could induce a bias in our results.

Assuming the spike in our data for higher number of bits precision formats is caused by a larger number of company affiliated accelerators in 2021, we can deduce that there

is an increase in higher precision formats over time, but it is most likely less than what is shown in Figure 13.

4.6. Neuromorphic Accelerators

Another interesting aspect that can be observed in the gathered data is the popularity of SNNs (Spiking Neural Networks) as an acceleration target. In Table 5, we list all low-power SNN accelerators gathered in this survey. Regarding the popularity of low-power SNN accelerators, we can observe from the table that it has not changed much over the four years, staying mostly the same. Even before 2019, many accelerators were designed for SNN acceleration. An example of this is IBM's TrueNorth from 2015 [16], which accelerates rate-based SNNs with a power consumption of around 70 mW. For more details on neuromorphic computing and hardware, we refer the reader to the surveys by Schuman et al. [21] and Shrestha et al. [20]. With regard to the power consumption, SNN accelerators tend to use much less power than its ANN-based counterparts. Calculating the mean and median power used in the accelerators in Table 5 results in a mean power of 0.9 W and a median power of 0.2 W for SNN accelerators, as compared to 1.7 W mean power and 0.5 W median power of non-SNN low-power accelerators.

Table 5. Low-power SNN accelerators.

Accelerator	Year	Tech	Area	Power	Freq.	Perf.	Power Eff.	Area Eff.	Type	Acc. Target
		nm	mm^2	mW	MHz	GOPS	GOPS/W	GOPS/mm^2		
Tianjic [61,62]	2019	28	14.44	950.0	300.0	1214.1 *	1278.0	84.1 *	ASIC	SNN, ANN
SpinalFlow [48]	2020	28	2.09	162.4	200.0	25.6 *†	157.6 *†	32.0 *†	simulated	SNN
NEBULA [51]	2020	32	86.73	5200.0	1200.0	-	-	-	simulated	SNN, ANN
YOSO [42]	2020	22	-	0.73	<1.0	1.3 *†	1757.8 *†	-	simulated	SNN
IMPULSE [28]	2021	65	0.09	0.2	200.0	0.2	990.0	2.2	ASIC	SNN
VSA [63]	2021	40	-	88.97	500.0	2304.0	25,900.0	-	simulated	SNN
Chen et al. [64]	2022	28	0.89	149.3	500.0	-	-	-	simulated	SNN
Skydiver [65]	2022	-	-	960.0	200.0	22.6 †	19.3 †	-	FPGA	SNN

* Calculated based on configuration and/or published metrics of specified accelerator. † Corresponding research paper display results in GSyOPS (Giga-Synaptic Operations per Second) instead of GOPS.

Looking closer at the acceleration targets in Table 5, two of the accelerators stand out from the rest, i.e., Tianjic [61,62] and NEBULA [51].

Tianjic [61,62] is a many-core AI research chip developed at Tsinghua University. It supports inference on SNN, ANN, and SNN-ANN hybrid models, where SNN-ANN hybrid models are models with mixed spiking (SNN) and non-spiking (ANN) layers. ANN, in this instant, refers to a variety of common neural network models (CNN, RNN, MLP), with arbitrary activation functions made possible with a look-up-table. Tianjic is able to accelerate more than one model at a time, due to its many-core architecture and decentralized structure.

NEBULA [51], developed at the Pennsylvania State University, is a spin-based, neuromorphic inference accelerator that accelerates SNN, ANN, and SNN-ANN hybrids, similar to Tianjic, barring the simultaneous execution of multiple models and the restriction on using ReLU as the only activation function. NEBULA accelerates one single network partitioned into spiking (SNN) and non-spiking (ANN) layers in the hybrid mode that preserves the low-power benefits of SNNs, with the additional advantage of reduced latency of SNNs over ANNs. NEBULA makes use of an ultra-low power spintronics-based Magnetic Tunnel Junction (MTJ) design for its neuron cores, instead of the traditional CMOS technology, to reduce the required voltage to be in the range of mV, instead of V.

Although popular in 2019 and 2020, the hybrid models have not gained traction in research in the later years, i.e., no more hybrid SNN-ANN accelerators were published in 2021 and 2022. Below, we go through some notable low-power SNN accelerators from the last four years.

YOSO [42] is a scalable and configurable low-power accelerator targeting TTFS-encoded SNNs. Time To First Spike (TTFS) is a common temporal encoding scheme in hardware SNN accelerators, but is usually not used in favor of the traditional rate-based scheme due to their low accuracy compared with the latter. YOSO solves this by introducing a new training algorithm that reduces the approximation error which accumulates as a result of converting ANNs to SNNs, thereby increasing the accuracy. Thus, TTFS-encoded SNNs can be considered for traditional ANN tasks, at higher efficiency and with comparable accuracy (within 0.2% to ANN). In addition, YOSO operates at very low clock frequencies, i.e., <1.0 MHz.

IMPULSE [28] is a low-power SNN accelerator, incorporating a SRAM-based compute-in-memory (CIM) macro for all necessary instructions in a SNN, including a fused weight and membrane potential memory. Further, IMPULSE consumes extremely low power, i.e., only 0.2 mW. The proposed SRAM-based CIM allows for a decreased memory access time compared to previous SNN accelerators; membrane potential usually incurs additional memory accesses in traditional SNN hardware.

VSA [65] is a configurable low-power SNN accelerator for both inference and training. As shown in Table 5, VSA has the highest power efficiency of the reviewed SNN accelerators. A binary weight spiking model which integrates and fires based on batch normalization is proposed, which allows for small time steps when direct training with the input encoding layer and spatio-temporal back propagation. The previous mentioned model and the support for different inference spike times and multi-bit input to the encoding layer, allowing VSA to have a high power efficiency.

4.7. Summary of Company Accelerators

This section is dedicated to accelerators where the authors are affiliated with a company. In Table 6, we have gathered all accelerators by different companies. One thing to notice in the table is the relatively high power of the company affiliated accelerators compared to the other accelerators discussed in this survey. With a mean power of 2.9 W and a median power of 1.9 W, the company affiliated accelerators have a 2.2x increase in mean power and 9.5x increase in median power compared to non-company affiliated accelerators (1.3 W and 0.2 W, respectively, for non-company affiliated accelerators). This could indicate that the power used in real hardware are higher than what the simulated systems usually predict. Below, we go through some selected company affiliated AI accelerators in more detail.

Table 6. Low-power AI accelerators by companies.

Company	Accelerator	Year	Tech	Area	Power	Freq.	Perf.	Power Eff.	Area Eff.	Type
			nm	mm^2	mW	MHz	GOPS	GOPS/W	GOPS/mm^2	
Samsung	NPU'19 [66]	2019	8	5.50	796.00	933.0	4423.5	8000.0	804.3 *	ASIC
Gyrfalcon †	Lightspeeur 5801S [59]	2019	28	36.00	224.00	200.0	2800.0	12,600.0	77.8 *	ASIC
Gyrfalcon †	Lightspeeur 2801S [60]	2019	-	49.00	600.00	100.0	5600.0	9300.0	114.3 *	ASIC
Hailo	Hailo-8 [67]	2019	-	225.00	2500.00	-	26,000.0	10,400.0 *	115.6 *	ASIC
Kneron	KL520 [57]	2019	-	-	575.00 *	300.0	345.0	600.0	-	ASIC
NVIDIA	Jetson Nano [33]	2019	-	3150.00	7500.00	-	250.0	37.5 *	0.1 *	ASIC
IBM	Oh et al. [56]	2020	14	9.80	2727.00 *	1500.0	3000.0	1100.0	300.0	ASIC
ARM	ETHOS-U55 [55]	2020	16	0.10	-	1000.0	512.0	-	2880.0 *	ASIC
ARM	ETHOS-U65 [68]	2020	16	0.60	-	-	1000.0	-	1666.7 *	ASIC
Kneron	KL720 [58]	2020	-	-	1362.50	696.0	1500.0	1250.0 *	-	ASIC
Cambricon	Cambricon-Q [69]	2021	45	9.18 *	1030.31	1000.0	2000.0	1941.2 *	217.9 *	simulated
IBM	RaPiD [70,71]	2021	7	36.00	6206.00 *	1600.0	51,200.0	8250.0	2610.0	ASIC
Samsung	NPU'21 [72,73]	2021	5	5.46	327.00	1196.0	14,700.0	13,600.0	4447.2 *	simulated
ARM	Ethos-N78 [74]	2021	-	-	-	-	10,000.0	-	-	ASIC
Kneron	KL530 [75]	2021	-	-	500.00	-	1000.0	2000.0 *	-	ASIC
quadric	q16 [76]	2021	16	256.00	5500.00	1000.0	64,000.0	11,636.4 *	253.9	ASIC
Mythic	M1076 AMP [77]	2021	-	294.50	4000.00	-	25,000.0	6250.0 *	84.9	ASIC
Xilinx	VE2002 [78]	2021	7	-	6000.00	-	5000.0	833.3 *	-	ASIC
Xilinx	VE2102 [78]	2021	7	-	7000.00	-	8000.0	1142.9 *	-	ASIC

* Calculated based on configuration and/or published metrics of specified accelerator. † Gyrfalcon Technology Inc.

RaPID [70] is a server/data center AI accelerator for ultra-low precision training and inference proposed by IBM researchers. RaPID accelerates CNNs, LSTMs, and Transformers. According to the authors, a single-core includes all necessary features to be used as a single-core edge accelerator. Thus, RaPID are designed for dual use in both data centers and edge devices.

NPU'21 [72] is a commercial neural processing unit (NPU) architecture for Samsung's flagship mobile system-on-chip (SoC). Using an energy-efficient inner product engine that utilizes the input feature map sparsity in combination with a re-configurable MAC array, a single NPU core achieves a throughput of 290.7 frames/s and 13.6 TOPS/W when executing an 8-bit quantized Inception-v3 model. NPU uses the latest technology (5 nm) and has the highest power efficiency of all reviewed company affiliated accelerators.

Gyrfalcon Technology's Lightspeeur 5801S [59] is an ultra low-power AI accelerator, targeting only CNN models. The accelerator is used in LG Electronics's smartphone as its AI chip. The accelerator's AI core is Gyrfalcon's patented APiM^{TM} (Al Processing in Memory) technology, allowing for on-chip parallelism. Lightspeeur 5801S has the lowest power of all reviewed company affiliated accelerators.

ARM's ETHOS-U55 [55] is a micro neural processing unit (NPU) for area-constrained embedded and IoT devices. It has a re-configurable MAC array, and can compute the core kernels used in many neural network tasks, such as convolutions, LSTM, RNN, pooling, activation functions, and primitive element wise functions, while other kernels run on the host CPU. ETHOS-U55 has the smallest area of all reviewed company affiliated accelerators.

Quadric's first generation AI accelerator, q16 [76], uses 256 Vortex cores (4 kB per core memory) with 8 MB on-chip memory. The accelerator supports multiple precision formats for its MAC operations and comes embedded into an M.2 form factor. q16 has the highest throughput of all reviewed company affiliated accelerators.

One observation that we made when going through the company accelerators was that none of them use fixed point arithmetic, which several research accelerators do. An explanation could be that when a company designs and implements an accelerator, both the development and the life time of the product are long. Thus, by using more flexible formats, such as floating point arithmetic, companies are more prepared for changes in future workloads.

4.8. Summary of Recent AI Accelerators

Summarizing the results from the previous five sections, we observe that company affiliated low-power AI accelerators tend to have a higher throughput ($\geq$10 GOPS) and power consumption ($\geq$100 mW) compared with non-company affiliated accelerators (2.2x less power consumption than company affiliated accelerators). This observation can be observed in other aspects of the accelerators too, i.e., higher overall frequency, better power efficiency, and smaller area. This could indicate that the power used in real hardware is higher than what the simulated systems usually predict. Another reason is the use of different technology parameters in simulations vs. real implementations, e.g., different circuit technology generations. In addition, the number of targets for acceleration is likely to contribute to the increase. Company affiliated accelerators tend to accelerate multiple targets, i.e., they accelerate applications which theoretically belong to the same general group of applications, sharing many internal components among them, compared with non-company affiliated accelerators.

We observed that integers continue to be the most common precision format used ($\leq$8-bit) for low-power AI accelerators. The fixed point (FxP) format and Google Brain's bfloat (BFP) format have increased in popularity during the last four years. We contribute the former to stricter power requirements, and the latter to the transition period from when it was announced (2018) to being used in AI accelerators in the low-power domain.

Finally, we have observed that low-power SNN accelerators have increased in popularity in recent years. A possible reason for that is that they tend to use less power than non-SNN accelerators (47% less power).

5. Discussion

As we presented in Section 4, simulations and ASIC dominate the field of research and commercial low-power AI accelerators between the years, 2019 to 2022, respectively. However, according to the study by Talib et al. [24], FPGA-based accelerators and CPU-based processors were mainstream between 2009 and 2019. This differs from our observations (excluding CPU-based processors, as we did not consider them in our survey) where FPGA-based accelerators are not common (4 of 79 reviewed accelerators are FPGA-based systems). In our survey, we have found that most company affiliated accelerators are ASICs, and not FPGA-based.

According to Reuther et al. [5,7,8], accelerators that have a power consumption below 100 W focuses solely on inference. However, our results show that is not the case. From our observations and findings, AI accelerators dedicated to both training and inference do exist in the low-power domain. Their power consumption tend to be in the higher spectrum of low-power ($\geq \approx$90 mW) for AI accelerators.

We also observed that ASIC-based accelerators have higher power requirements compared to simulated systems. Our results show a clear distinction between the two groups (simulated and ASIC). This could be caused by the simulation itself; as we cannot be certain of how accurate the simulations models real hardware. When simulating an accelerator based on an architecture, it is common to use systems that model the hardware by specifying parameters for the energy consumption of each component. Therefore, the resulting power consumption for the simulated system is based on the values chosen to represent the energy consumption of the components, and this will induce variations in the consumption of energy compared to a real hardware with physical components and fluctuating energy usage. Further, we have observed that real hardware implementations (by companies) use more advanced circuit technology to increase performance, thus increasing clock frequencies, which in turn drive a higher power consumption.

Regarding the numerical precision formats, we found that the 8-bit and 16-bit formats are still prevalently used in AI accelerators, even more so in the low-power domain. We observed that the fixed point and bfloat formats have increased in usage in non-company low-power accelerators, and we speculate that the increase in fixed point is due to the stricter power requirements and the increase in the usage of the bfloat is due to the transition period from when it was announced until it is used in the low-power domain. Our argument is two-fold: (i) the fact that research in low-power AI accelerators tend to focus on gaining higher throughput by reducing flexibility and as the fixed point format locks and restricts the number of decimals used, and (ii) as the bfloat format uses a wider dynamic range of values, but with reduced accuracy, they are often preferred over the traditional floating point formats.

We observed an increased interest in Transformers, see Section 4.4. Our intuition for this increase is based on the fact that the Transformer is a recent algorithm (published in 2017) and therefore the increased interest now is due to the transition time into the low-power domain for Transformer models. Since Transformer models in general are very computational demanding, the transition into the low-power domain occurred later (2021). We predict that the popularity of the Transformer will continue in the coming years.

In Section 4.4, we observed that accelerators could accelerate more than one acceleration target, and when that is the case, the targets for acceleration are often applications that are similar to each other. We found that this is most prevalent in company affiliated accelerators; where they tend to support multiple applications, meaning they can accelerate multiple acceleration targets. Assuming this will continue to be the case for later years, and that the company affiliated accelerators continue to trade higher throughput for higher power (to be able to achieve the former results), we can assume that the lowest power that a commercial low-power AI accelerator will follow the current trend, i.e., a power consumption higher than $\approx$200 mW.

6. Conclusions

In this paper, we present an overview of the recent development of low-power AI accelerators for the last four years. In total, we have covered 79 different accelerators, both from academia and industry. The number of accelerators presented has been quite stable in between the years, but a majority of the accelerators seem to come from research organizations. It seems that very low-power (<100 mW) accelerators are quite rare, as are devices with a very high throughput (>10 TOPS/W). However, it should be noted that the trend regarding power efficiency seems to follow the trends observed in previous, non-low-power accelerators [8]. Low-power accelerators are within the same trend-lines, only with a lower power consumption.

Interestingly, there does not seem to be any increase in power efficiency over the years among the observed accelerators. Instead, most accelerators seems to cluster between 100 mW–10 W and between 100 GOPS/W–10 TOPS/W. Another interesting aspect to note is that while companies implement more accelerators (i.e., non-simulated), the power consumption for company accelerators tends to be higher compared to non-companies (mean power of 2.9 W compared to 1.3 W for non-company accelerators). This might be a consequence of how power consumption is calculated for simulated devices, which could be considered a theoretical lower limit. It could also be because implemented company accelerators tend to accelerate multiple targets.

Finally, while CNNs and DNNs are the most popular accelerator targets, Transformers are becoming more popular. While interest in SNN seems to be quite stable, spiking accelerators tend to use less power than non-spiking (mean power of 0.9 W and 1.7 W, respectively). As such, it is an interesting technology for low-power environments.

Author Contributions: Conceptualization, C.Å., H.G. and A.B.; methodology, C.Å., H.G. and A.B.; validation, C.Å., H.G. and A.B.; investigation, C.Å.; data curation, C.Å.; writing—original draft preparation, C.Å.; writing—review and editing, C.Å., H.G. and A.B.; visualization, C.Å.; supervision, H.G. and A.B.; project administration, H.G.; funding acquisition, H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Excellence Center at Linköping-Lund in Information Technology (ELLIIT) project "GPAI—General Purpose AI Computing" and by the Knowledge-foundation in Sweden through the project "Rekrytering17, Lektor i datavetenskap" (grant 20170236).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data collected for this paper is available at https://github.com/caleskog/Low-Power-AI-Accelerators-Data, accessed on 28 September 2022.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

All accelerators included in this survey are listed in Tables A1–A4. Due to size restrictions, some information are omitted.

Table A1. Low-power AI accelerators from 2019.

Accelerator	Acc. Target	Tech	Area	Power	Freq.	Perf.	Power Eff.	Area Eff.	Type
		nm	mm^2	mW	MHz	GOPS	GOPS/W	$GOPS/mm^2$	
Eyeriss v2 [47]	CNN	65	-	476.25 *	200.0	153.6	578.3	-	simulated
Khabbazan et al. [79]	CNN	-	-	1770.00	160.0	41.0	23.1	-	FPGA
Wang et al. [31]	CNN	40	2.16	68.00	50.0	3.2 *	47.1 *	1.5 *	FPGA
SparTen [80]	CNN	45	0.77	118.30	800.0	25.6 *	216.4 *	33.4 *	simulated
DT-CNN [81]	CNN, GAN	65	6.80	196.00	200.0	639.7	3260.0	94.1	simulated
Dante [82]	DNN	14	2.30	-	330.0	-	-	-	simulated
TIE [83]	DNN	28	1.74	154.80	1000.0	7640.0	72,900.0	43,807.3 *	simulated
MnnFast [84]	Transformer	28	-	-	1000.0	120.0	-	-	FPGA
ExTensor [85]	SpGEMM	32	98.89	-	1000.0	128.0	-	-	simulated
WAX [29]	CNN	28	0.33	1.55 *	200.0	26.2 *	16,938.5 *	206.0	simulated
CGNet-NM [41]	CNN	28	3.86	159.50	500.0	144.0 *	902.8 *	37.3 *	simulated
Tianjic [61,62]	SNN, ANN	28	14.44	950.00	300.0	1214.1 *	1278.0	84.1 *	ASIC
MASR [86]	RNN	16	4.70	375.00	1000.0	544.0 *	1051.4 *	194.6 *	simulated
Li et al. [87]	CNN	28	10.92	154.03	240.0	604.7	2970.0	55.4 *	ASIC
NPU'19 [66] ‡	DNN, CNN	8	5.50	796.00	933.0	4423.5	8000.0	804.3 *	ASIC
LNPU [88]	CNN, LSTM, RNN	45	16.00	367.00	200.0	2482.8 *	3382.5 *	77.6 *	ASIC
eCNN [89]	CNN	40	55.23	6950.00	250.0	41,000.0	5899.3	742.4	simulated
Lightspeeur 5801S [59] ‡	CNN	28	36.00	224.00	200.0	2800.0	12,600.0	77.8 *	ASIC
Lightspeeur 2801S [60] ‡	CNN	-	49.00	600.00	100.0	5600.0	9300.0	114.3 *	ASIC
Hailo-8 [67] ‡	DNN	-	225.00	2500.00	-	26,000.0	10,400.0 *	115.6 *	ASIC
KL520 [57] ‡	LSTM, CNN	-	-	575.00 *	300.0	345.0	600.0	-	ASIC
Jetson Nano [33] ‡	DNN	-	3150.00	7500.00	-	250.0	37.5 *	0.1 *	ASIC

* Calculated based on configuration and/or published metrics of specified accelerator. ‡ Published research paper's authors are affiliated with a company (see Table 6 for specifics).

Table A2. Low-power AI accelerators from 2020.

Accelerator	Acc. Target	Tech	Area	Power	Freq.	Perf.	Power Eff.	Area Eff.	Type
		nm	mm^2	mW	MHz	GOPS	GOPS/W	$GOPS/mm^2$	
JPEG-ACT [90]	CNN	-	1.48	1360.00	1455.0	-	-	-	simulated
TIMELY [32]	DNN, CNN	65	91.00	0.41 *	40.0	4.2 *	21000.0	0.0 *	simulated
RecNMP [54]	PR	40	0.54	169.25	240.0	-	-	-	simulated
uGEMM [91]	GEMM	45	0.60	165.00	400.0	102.4 *	670.5 *	185.6 *	simulated
SpinalFlow [48]	SNN	28	2.09	162.40	200.0	25.6 *†	157.6 *†	32.0 *†	simulated
NEBULA [51]	SNN, ANN	32	86.73	5200.00	1200.0	-	-	-	simulated
XpulpNN [43]	DNN, CNN	22	1.00	100.00	250.0	5.0	550.0	55.0 *	simulated
YOSO [42]	SNN	22	-	0.73	<1.0	1.3 *†	1757.8 *†	-	simulated
DRQ [44]	DNN	45	0.32	71.96 *	500.0	1584.0 *	22160.2 *	4950.0 *	simulated
SmartExchange [92]	DNN, CNN	28	-	2853.00 *	1000.0	8192.0 *	4054.0 *	-	simulated
MatRaptor [93]	SpGEMM	28	2.26	1344.95	2000.0	32.0	23.8 *	14.2 *	simulated
DT-CNN [94]	CNN, GAN	65	6.80	206.00	200.0	664.4	3220.0	97.6	simulated
Zhang et al. [35]	MAC	45	-	199.68	25.0	121.4	610.0	-	simulated
A3 [95]	Transformer	40	2.08	98.92	1000.0	221.0	2234.1 *	106.1 *	simulated
OPTIMUS [96]	GEMM	28	5.19	928.14	200.0	-	-	-	simulated
SpArch [97]	SpGEMM	40	28.49	9260.00	1000.0	16.0	1.7 *	0.6 *	simulated
Oh et al. [56] ‡	CNN, LSTM, RNN	14	9.80	2727.00 *	1500.0	3000.0	1100.0	300.0	ASIC
REVEL [98]	Matrix Alg.	28	1.93	1663.30	1250.0	-	-	-	simulated
ETHOS-U55 [55] ‡	CNN, LSTM, RNN	16	0.10	-	1000.0	512.0	-	2880.0 *	ASIC
ETHOS-U65 [68] ‡	CNN, LSTM, RNN	16	0.60	-	-	1000.0	-	1666.7 *	ASIC
KL720 [58] ‡	CNN, LSTM, RNN	-	-	1362.50	696.0	1500.0	1250.0 *	-	ASIC

* Calculated based on configuration and/or published metrics of specified accelerator. † Corresponding research paper display results in GSyOPS (Giga-Synaptic Operations per Second) instead of GOPS. ‡ Published research paper's authors are affiliated with a company (see Table 6 for specifics).

Table A3. Low-power AI accelerators from 2021.

Accelerator	Acc. Target	Tech	Area	Power	Freq.	Perf.	Power Eff.	Area Eff.	Type
		nm	mm^2	mW	MHz	GOPS	GOPS/W	$GOPS/mm^2$	
GoSPA [30]	CNN	28	0.25 *	26.64	500.0	16.0 *	600.6 *	64.0 *	simulated
RingCNN [45]	CNN	40	23.36	2220.00	250.0	41.0	28,400.0	1755.1 *	simulated
SNAFU-ARCH [6]	GEMM	28	1.00	0.32	50.0	0.2 *	625.0 *	0.2 *	simulated
Cambricon-Q [69] ‡	CNN, RNN	45	9.18 *	1030.31	1000.0	2000.0	1941.2 *	217.9 *	simulated
ELSA [99]	Transformer	40	2.10	1472.89	1000.0	1088.0	738.7 *	518.1 *	simulated
RaPiD [70,71] ‡	CNN, LSTM, RNN, Transformer	7	36.00	6206.00 *	1600.0	51,200.0	8250.0	2610.0	ASIC
NPU'21 [72,73] ‡	DNN, CNN	5	5.46	327.00	1196.0	14,700.0	13,600.0	4447.2 *	simulated
Lin et al. [37]	MLP	40	0.57	93.00	1600.0	899.0	9700.0	1577.2 *	simulated
PremCNN [46]	DNN, CNN	28	3.44	500.00	800.0	102.4 *	204.8 *	29.8 *	simulated
GAMMA [100]	SpGEMM	45	30.60	-	1000.0	20.0	-	0.7 *	simulated
FPRaker [38]	CNN	65	771.61 *	3942.00 *	600.0	11,059.2 *	2835.7 *	14.3 *	simulated
Bitlet [101]	DNN, CNN, GAN, Transformer	28	1.54	366.00	1000.0	744.7	1335.9	483.6 *	simulated
RASA [102]	GEMM	15	0.90 *	-	500.0	128.0 *	-	138.5 *	simulated
StepStone PIM [103]	GEMM	-	-	1400.00	1200.0	-	-	-	simulated
SpAtten [39]	Transformer	40	18.71	8300.00	1000.0	2880.0 *	347.0 *	153.9 *	simulated
IMPULSE [28]	SNN	65	0.09	0.20	200.0	0.2	990.0	2.2	ASIC
VSA [63]	SNN	40	-	88.97	500.0	2304.0	25900.0	-	simulated
Zeng et al. [104]	CNN	65	-	478.00	250.0	576.0	1205.0	-	simulated
Ethos-N78 [74] ‡	CNN, LSTM, RNN	-	-	-	-	10,000.0	-	-	ASIC
KL530 [75] ‡	CNN, LSTM, RNN, Transformer	-	-	500.00	-	1000.0	2000.0 *	-	ASIC
q16 [76] ‡	DNN	16	256.00	5500.00	1000.0	64,000.0	11636.4 *	253.9	ASIC
M1076 AMP [77] ‡	DNN	-	294.50	4000.00	-	25,000.0	6250.0 *	84.9	ASIC
VE2002 [78] ‡	DNN, CNN	7	-	6000.00	-	5000.0	833.3*	-	ASIC
VE2102 [78] ‡	DNN, CNN	7	-	7000.00	-	8000.0	1142.9 *	-	ASIC

* Calculated based on configuration and/or published metrics of specified accelerator. ‡ Published research paper's authors are affiliated with a company (see Table 6 for specifics).

Table A4. Low-power AI accelerators from 2022.

Accelerator	Acc. Target	Tech	Area	Power	Freq.	Perf.	Power Eff.	Area Eff.	Type
		nm	mm^2	mW	MHz	GOPS	GOPS/W	$GOPS/mm^2$	
RedMulE [27]	GEMM	22	0.50	43.50	476.0	30.0	688.0	60.0 *	simulated
DOTA [36]	Transformer	22	2.75	3020.00	100.0	2048.0 *	678.1 *	745.8 *	simulated
EcoFlow [49]	CNN, GAN	45	-	332.00	200.0	39.0 *	117.5 *	-	simulated
FINGERS [105]	Graph Mining	28	18.68 *	3682.00 *	1000.0	-	-	-	simulated
DTQAtten [106]	Transformer	40	1.41	-	1000.0	952.8	-	678.4	simulated
Chen et al. [64]	SNN	28	0.89	149.30	500.0	-	-	-	simulated
Skydiver [65]	SNN	-	-	960.00	200.0	22.6 †	19.3 †	-	FPGA
DIMMining [40]	Graph Mining	32	0.38	105.82	500.0	-	-	-	simulated
MeNDA-PU [50]	SpMatrix Transposition	40	7.10	92.40	800.0	-	-	-	simulated
Ubrain [107]	DNN, CNN, RNN	32	4.00	45.00	400.0	-	-	-	simulated
Mokey [108]	Transformer	65	23.90	4478.00 *	1000.0	-	-	-	simulated
HP-LeOPArd [109]	Attention	65	3.47	-	800.0	574.1	-	165.5	simulated

* Calculated based on configuration and/or published metrics of specified accelerator. † Corresponding research paper display results in GSyOPS (Giga-Synaptic Operations per Second) instead of GOPS.

References

1. Amant, R.S.; Jiménez, D.A.; Burger, D. Low-power, high-performance analog neural branch prediction. In Proceedings of the 2008 41st IEEE/ACM International Symposium on Microarchitecture, Lake Como, Italy, 8–12 November 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 447–458.
2. Chen, Y.; Zheng, B.; Zhang, Z.; Wang, Q.; Shen, C.; Zhang, Q. Deep Learning on Mobile and Embedded Devices: State-of-the-Art, Challenges, and Future Directions. *ACM Comput. Surv.* **2020**, *53*, 1–37. [CrossRef]
3. Theis, T.N.; Wong, H.S.P. The End of Moore's Law: A New Beginning for Information Technology. *Comput. Sci. Eng.* **2017**, *19*, 41–50. [CrossRef]
4. Hennessy, J.L.; Patterson, D.A. A New Golden Age for Computer Architecture. *Commun. ACM* **2019**, *62*, 48–60. [CrossRef]

5. Reuther, A.; Michaleas, P.; Jones, M.; Gadepally, V.; Samsi, S.; Kepner, J. Survey and Benchmarking of Machine Learning Accelerators. In Proceedings of the 2019 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 24–26 September 2019; pp. 1–9. [CrossRef]
6. Gobieski, G.; Atli, A.O.; Mai, K.; Lucia, B.; Beckmann, N. Snafu: An Ultra-Low-Power, Energy-Minimal CGRA-Generation Framework and Architecture. In Proceedings of the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 14–18 June 2021; pp. 1027–1040. [CrossRef]
7. Reuther, A.; Michaleas, P.; Jones, M.; Gadepally, V.; Samsi, S.; Kepner, J. Survey of Machine Learning Accelerators. In Proceedings of the 2020 IEEE High Performance Extreme Computing Conference (HPEC), Greater Boston Area, MA, USA, 22–24 September 2020; pp. 1–12. [CrossRef]
8. Reuther, A.; Michaleas, P.; Jones, M.; Gadepally, V.; Samsi, S.; Kepner, J. AI Accelerator Survey and Trends. In Proceedings of the 2021 IEEE High Performance Extreme Computing Conference (HPEC), Virtual, 19–23 September 2022; pp. 1–9. [CrossRef]
9. Lin, W.; Adetomi, A.; Arslan, T. Low-Power Ultra-Small Edge AI Accelerators for Image Recognition with Convolution Neural Networks: Analysis and Future Directions. *Electronics* **2021**, *10*, 2048. [CrossRef]
10. Sze, V.; Chen, Y.H.; Yang, T.J.; Emer, J.S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proc. IEEE* **2017**, *105*, 2295–2329. [CrossRef]
11. Nabavinejad, S.M.; Baharloo, M.; Chen, K.C.; Palesi, M.; Kogel, T.; Ebrahimi, M. An Overview of Efficient Interconnection Networks for Deep Neural Network Accelerators. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2020**, *10*, 268–282. [CrossRef]
12. Chen, T.; Du, Z.; Sun, N.; Wang, J.; Wu, C.; Chen, Y.; Temam, O. DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning. In Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems, Salt Lake City, UT, USA, 1–5 March 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 269–284. [CrossRef]
13. Chen, Y.; Luo, T.; Liu, S.; Zhang, S.; He, L.; Wang, J.; Li, L.; Chen, T.; Xu, Z.; Sun, N.; et al. DaDianNao: A Machine-Learning Supercomputer. In Proceedings of the 2014 47th Annual IEEE/ACM International Symposium on Microarchitecture, Cambridge, UK, 19–17 December 2014; pp. 609–622. [CrossRef]
14. Du, Z.; Fasthuber, R.; Chen, T.; Ienne, P.; Li, L.; Luo, T.; Feng, X.; Chen, Y.; Temam, O. ShiDianNao: Shifting Vision Processing Closer to the Sensor. *SIGARCH Comput. Archit. News* **2015**, *43*, 92–104. [CrossRef]
15. Liu, D.; Chen, T.; Liu, S.; Zhou, J.; Zhou, S.; Teman, O.; Feng, X.; Zhou, X.; Chen, Y. PuDianNao: A Polyvalent Machine Learning Accelerator. In Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems, Vancouver, Canada, 25–29 March 2023; Association for Computing Machinery: New York, NY, USA, 2015; pp. 369–381. [CrossRef]
16. Akopyan, F.; Sawada, J.; Cassidy, A.; Alvarez-Icaza, R.; Arthur, J.; Merolla, P.; Imam, N.; Nakamura, Y.; Datta, P.; Nam, G.J.; et al. TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip. *IEEE Trans.-Comput.-Aided Des. Integr. Circuits Syst.* **2015**, *34*, 1537–1557. [CrossRef]
17. DeBole, M.V.; Taba, B.; Amir, A.; Akopyan, F.; Andreopoulos, A.; Risk, W.P.; Kusnitz, J.; Ortega Otero, C.; Nayak, T.K.; Appuswamy, R.; et al. TrueNorth: Accelerating From Zero to 64 Million Neurons in 10 Years. *Computer* **2019**, *52*, 20–29. [CrossRef]
18. Chen, Y.H.; Krishna, T.; Emer, J.S.; Sze, V. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE J. Solid-State Circuits* **2017**, *52*, 127–138. [CrossRef]
19. Ibtesam, M.; Solangi, U.S.; Kim, J.; Ansari, M.A.; Park, S. Highly Efficient Test Architecture for Low-Power AI Accelerators. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2022**, *41*, 2728–2738. [CrossRef]
20. Shrestha, A.; Fang, H.; Mei, Z.; Rider, D.P.; Wu, Q.; Qiu, Q. A Survey on Neuromorphic Computing: Models and Hardware. *IEEE Circuits Syst. Mag.* **2022**, *22*, 6–35. [CrossRef]
21. Schuman, C.D.; Potok, T.E.; Patton, R.M.; Birdwell, J.D.; Dean, M.E.; Rose, G.S.; Plank, J.S. A Survey of Neuromorphic Computing and Neural Networks in Hardware. *arXiv* **2017**, arXiv:1705.06963.
22. Seo, J.s.; Saikia, J.; Meng, J.; He, W.; Suh, H.s.; Anupreetham; Liao, Y.; Hasssan, A.; Yeo, I. Digital Versus Analog Artificial Intelligence Accelerators: Advances, trends, and emerging designs. *IEEE-Solid-State Circuits Mag.* **2022**, *14*, 65–79. [CrossRef]
23. Sunny, F.P.; Taheri, E.; Nikdast, M.; Pasricha, S. A Survey on Silicon Photonics for Deep Learning. *J. Emerg. Technol. Comput. Syst.* **2021**, *17*, 1–57. [CrossRef]
24. Talib, M.A.; Majzoub, S.; Nasir, Q.; Jamal, D. A systematic literature review on hardware implementation of artificial intelligence algorithms. *J. Supercomput.* **2021**, *77*, 1897–1938. [CrossRef]
25. Li, W.; Liewig, M. A survey of AI accelerators for edge environment. In Proceedings of the World Conference on Information Systems and Technologies, Budva, Montenegro, 7–10 April 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 35–44.
26. Wohlin, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, London, UK, 13–14 May 2014; pp. 1–10.
27. Tortorella, Y.; Bertaccini, L.; Rossi, D.; Benini, L.; Conti, F. RedMulE: A Compact FP16 Matrix-Multiplication Accelerator for Adaptive Deep Learning on RISC-V-Based Ultra-Low-Power SoCs. *arXiv* **2022**, arXiv:2204.11192.
28. Agrawal, A.; Ali, M.; Koo, M.; Rathi, N.; Jaiswal, A.; Roy, K. IMPULSE: A 65-nm Digital Compute-in-Memory Macro With Fused Weights and Membrane Potential for Spike-Based Sequential Learning Tasks. *IEEE-Solid-State Circuits Lett.* **2021**, *4*, 137–140. [CrossRef]

29. Gudaparthi, S.; Narayanan, S.; Balasubramonian, R.; Giacomin, E.; Kambalasubramanyam, H.; Gaillardon, P.E. Wire-Aware Architecture and Dataflow for CNN Accelerators. In Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, Columbus, OH, USA, 12–16 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–13. [CrossRef]
30. Deng, C.; Sui, Y.; Liao, S.; Qian, X.; Yuan, B. GoSPA: An Energy-efficient High-performance Globally Optimized SParse Convolutional Neural Network Accelerator. In Proceedings of the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 14–18 June 2021; pp. 1110–1123. [CrossRef]
31. Wang, M.; Chandrakasan, A.P. Flexible Low Power CNN Accelerator for Edge Computing with Weight Tuning. In Proceedings of the 2019 IEEE Asian Solid-State Circuits Conference (A-SSCC), Macau, China, 4–6 November 2019; pp. 209–212. [CrossRef]
32. Li, W.; Xu, P.; Zhao, Y.; Li, H.; Xie, Y.; Lin, Y. Timely: Pushing Data Movements And Interfaces In Pim Accelerators Towards Local And In Time Domain. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Virtual, 30 May–3 June 2020; pp. 832–845. [CrossRef]
33. NVIDIA Corporation. JETSON NANO. Available online: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-nano/product-development/ (accessed on 29 June 2022).
34. Dennard, R.; Gaensslen, F.; Yu, H.N.; Rideout, V.; Bassous, E.; LeBlanc, A. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE J.-Solid-State Circuits* **1974**, *9*, 256–268. [CrossRef]
35. Zhang, S.; Huang, K.; Shen, H. A Robust 8-Bit Non-Volatile Computing-in-Memory Core for Low-Power Parallel MAC Operations. *IEEE Trans. Circuits Syst. Regul. Pap.* **2020**, *67*, 1867–1880. [CrossRef]
36. Qu, Z.; Liu, L.; Tu, F.; Chen, Z.; Ding, Y.; Xie, Y. DOTA: Detect and Omit Weak Attentions for Scalable Transformer Acceleration. In Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Vancouver, Canada, 25–29 March 2023; Association for Computing Machinery: New York, NY, USA, 2022; pp. 14–26. [CrossRef]
37. Lin, W.C.; Chang, Y.C.; Huang, J.D. An Efficient and Low-Power MLP Accelerator Architecture Supporting Structured Pruning, Sparse Activations and Asymmetric Quantization for Edge Computing. In Proceedings of the 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS), Washington, DC, USA, 6–9 June 2021; pp. 1–5. [CrossRef]
38. Awad, O.M.; Mahmoud, M.; Edo, I.; Zadeh, A.H.; Bannon, C.; Jayarajan, A.; Pekhimenko, G.; Moshovos, A. FPRaker: A Processing Element For Accelerating Neural Network Training. In Proceedings of the MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture, Virtual, 18–22 October 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 857–869. [CrossRef]
39. Wang, H.; Zhang, Z.; Han, S. SpAtten: Efficient Sparse Attention Architecture with Cascade Token and Head Pruning. In Proceedings of the 2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA), Seoul, Korea, 27 February–3 March 2021; pp. 97–110. [CrossRef]
40. Dai, G.; Zhu, Z.; Fu, T.; Wei, C.; Wang, B.; Li, X.; Xie, Y.; Yang, H.; Wang, Y. DIMMining: Pruning-Efficient and Parallel Graph Mining on near-Memory-Computing. In Proceedings of the 49th Annual International Symposium on Computer Architecture, New York, NY, USA, 18–22 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 130–145. [CrossRef]
41. Hua, W.; Zhou, Y.; De Sa, C.; Zhang, Z.; Suh, G.E. Boosting the Performance of CNN Accelerators with Dynamic Fine-Grained Channel Gating. In Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, Columbus, OH, USA, 12–16 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 139–150. [CrossRef]
42. P, S.; Chu, K.T.N.; Tavva, Y.; Wu, J.; Zhang, M.; Li, H.; Carlson, T.E. You Only Spike Once: Improving Energy-Efficient Neuromorphic Inference to ANN-Level Accuracy. *arXiv* **2020**, arXiv:2006.09982.
43. Garofalo, A.; Tagliavini, G.; Conti, F.; Rossi, D.; Benini, L. XpulpNN: Accelerating Quantized Neural Networks on RISC-V Processors Through ISA Extensions. In Proceedings of the 2020 Design, Automation Test in Europe Conference Exhibition (DATE), Grenoble, France, 9–13 March 2020; pp. 186–191. [CrossRef]
44. Song, Z.; Fu, B.; Wu, F.; Jiang, Z.; Jiang, L.; Jing, N.; Liang, X. DRQ: Dynamic Region-based Quantization for Deep Neural Network Acceleration. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Virtual, 30 May–3 June 2020; pp. 1010–1021. [CrossRef]
45. Huang, C.T. RingCNN: Exploiting Algebraically-Sparse Ring Tensors for Energy-Efficient CNN-Based Computational Imaging. In Proceedings of the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 14–18 June 2021; pp. 1096–1109. [CrossRef]
46. Deng, C.; Liao, S.; Yuan, B. PermCNN: Energy-Efficient Convolutional Neural Network Hardware Architecture With Permuted Diagonal Structure. *IEEE Trans. Comput.* **2021**, *70*, 163–173. [CrossRef]
47. Chen, Y.H.; Yang, T.J.; Emer, J.; Sze, V. Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2019**, *9*, 292–308. [CrossRef]
48. Narayanan, S.; Taht, K.; Balasubramonian, R.; Giacomin, E.; Gaillardon, P.E. SpinalFlow: An Architecture and Dataflow Tailored for Spiking Neural Networks. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Virtual, 30 May–3 June 2020; pp. 349–362. [CrossRef]
49. Orosa, L.; Koppula, S.; Umuroglu, Y.; Kanellopoulos, K.; Gómez-Luna, J.; Blott, M.; Vissers, K.A.; Mutlu, O. EcoFlow: Efficient Convolutional Dataflows for Low-Power Neural Network Accelerators. *arXiv* **2022**, arXiv:2202.02310.

50. Feng, S.; He, X.; Chen, K.Y.; Ke, L.; Zhang, X.; Blaauw, D.; Mudge, T.; Dreslinski, R. MeNDA: A near-Memory Multi-Way Merge Solution for Sparse Transposition and Dataflows. In Proceedings of the 49th Annual International Symposium on Computer Architecture, New York, NY, USA, 18–22 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 245–258. [CrossRef]
51. Singh, S.; Sarma, A.; Jao, N.; Pattnaik, A.; Lu, S.; Yang, K.; Sengupta, A.; Narayanan, V.; Das, C.R. NEBULA: A Neuromorphic Spin-Based Ultra-Low Power Architecture for SNNs and ANNs. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Virtual, 30 May–3 June 2020; pp. 363–376. [CrossRef]
52. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
53. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
54. Ke, L.; Gupta, U.; Cho, B.Y.; Brooks, D.; Chandra, V.; Diril, U.; Firoozshahian, A.; Hazelwood, K.; Jia, B.; Lee, H.H.S.; et al. RecNMP: Accelerating Personalized Recommendation with Near-Memory Processing. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Virtual, 30 May–3 June 2020; pp. 790–803. [CrossRef]
55. ARM Limited. ARM MICRONPU ETHOS-U55. Available online: https://www.arm.com/products/silicon-ip-cpu/ethos/ethos-u55 (accessed on 29 June 2022).
56. Oh, J.; Lee, S.K.; Kang, M.; Ziegler, M.; Silberman, J.; Agrawal, A.; Venkataramani, S.; Fleischer, B.; Guillorn, M.; Choi, J.; et al. A 3.0 TFLOPS 0.62V Scalable Processor Core for High Compute Utilization AI Training and Inference. In Proceedings of the 2020 IEEE Symposium on VLSI Circuits, Honolulu, HI, USA, 14–19 June 2020; pp. 1–2. [CrossRef]
57. Kneron. KL520 AI SoC. Available online: https://www.kneron.com/cn/page/soc/ (accessed on 29 June 2022).
58. Kneron. KL720 AI SoC. Available online: https://www.kneron.com/cn/page/soc/ (accessed on 29 June 2022).
59. Gyrfalcon Technology Inc. LIGHTSPEEUR® 5801S NEURAL ACCELERATOR. Available online: https://www.gyrfalcontech.ai/solutions/lightspeeur-5801/ (accessed on 29 June 2022).
60. Gyrfalcon Technology Inc. LIGHTSPEEUR® 2801S NEURAL ACCELERATOR. Available online: https://www.gyrfalcontech.ai/solutions/2801s/ (accessed on 29 June 2022).
61. Pei, J.; Deng, L.; Song, S.; Zhao, M.; Zhang, Y.; Wu, S.; Wang, G.; Zou, Z.; Wu, Z.; He, W.; et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature* **2019**, *572*, 106–111.
62. Deng, L.; Wang, G.; Li, G.; Li, S.; Liang, L.; Zhu, M.; Wu, Y.; Yang, Z.; Zou, Z.; Pei, J.; et al. Tianjic: A Unified and Scalable Chip Bridging Spike-Based and Continuous Neural Computation. *IEEE J. -Solid-State Circuits* **2020**, *55*, 2228–2246. [CrossRef]
63. Lien, H.H.; Hsu, C.W.; Chang, T.S. VSA: Reconfigurable Vectorwise Spiking Neural Network Accelerator. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, 22–28 May 2021; pp. 1–5. [CrossRef]
64. Chen, Q.; He, G.; Wang, X.; Xu, J.; Shen, S.; Chen, H.; Fu, Y.; Li, L. A 67.5μJ/Prediction Accelerator for Spiking Neural Networks in Image Segmentation. *IEEE Trans. Circuits Syst. Ii Express Briefs* **2022**, *69*, 574–578. [CrossRef]
65. Chen, Q.; Gao, C.; Fang, X.; Luan, H. Skydiver: A Spiking Neural Network Accelerator Exploiting Spatio-Temporal Workload Balance. *IEEE Trans.-Comput.-Aided Des. Integr. Circuits Syst.* **2022**, 1. [CrossRef]
66. Song, J.; Cho, Y.; Park, J.S.; Jang, J.W.; Lee, S.; Song, J.H.; Lee, J.G.; Kang, I. 7.1 An 11.5TOPS/W 1024-MAC Butterfly Structure Dual-Core Sparsity-Aware Neural Processing Unit in 8nm Flagship Mobile SoC. In Proceedings of the 2019 IEEE International Solid- State Circuits Conference—(ISSCC), San Francisco, CA, USA, 17–21 February 2019; pp. 130–132. [CrossRef]
67. Hailo. Hailo-8™ AI Processor. Available online: https://hailo.ai/product-hailo/hailo-8 (accessed on 29 June 2022).
68. ARM Limited. ARM MICRONPU ETHOS-U65. Available online: https://www.arm.com/products/silicon-ip-cpu/ethos/ethos-u65 (accessed on 29 June 2022).
69. Zhao, Y.; Liu, C.; Du, Z.; Guo, Q.; Hu, X.; Zhuang, Y.; Zhang, Z.; Song, X.; Li, W.; Zhang, X.; et al. Cambricon-Q: A Hybrid Architecture for Efficient Training. In Proceedings of the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 14–18 June 2021; pp. 706–719. [CrossRef]
70. Venkataramani, S.; Srinivasan, V.; Wang, W.; Sen, S.; Zhang, J.; Agrawal, A.; Kar, M.; Jain, S.; Mannari, A.; Tran, H.; et al. RaPiD: AI Accelerator for Ultra-low Precision Training and Inference. In Proceedings of the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 14–18 June 2021; pp. 153–166. [CrossRef]
71. Agrawal, A.; Lee, S.K.; Silberman, J.; Ziegler, M.; Kang, M.; Venkataramani, S.; Cao, N.; Fleischer, B.; Guillorn, M.; Cohen, M.; et al. 9.1 A 7nm 4-Core AI Chip with 25.6TFLOPS Hybrid FP8 Training, 102.4TOPS INT4 Inference and Workload-Aware Throttling. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; Volume 64, pp. 144–146. [CrossRef]
72. Jang, J.W.; Lee, S.; Kim, D.; Park, H.; Ardestani, A.S.; Choi, Y.; Kim, C.; Kim, Y.; Yu, H.; Abdel-Aziz, H.; et al. Sparsity-Aware and Re-configurable NPU Architecture for Samsung Flagship Mobile SoC. In Proceedings of the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 14–18 June 2021; pp. 15–28. [CrossRef]
73. Park, J.S.; Jang, J.W.; Lee, H.; Lee, D.; Lee, S.; Jung, H.; Lee, S.; Kwon, S.; Jeong, K.; Song, J.H.; et al. 9.5 A 6K-MAC Feature-Map-Sparsity-Aware Neural Processing Unit in 5nm Flagship Mobile SoC. In Proceedings of the 2021 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 13–22 February 2021; Volume 64, pp. 152–154. [CrossRef]

74. ARM Limited. ARM NPU ETHOS-N78. Available online: https://www.arm.com/products/silicon-ip-cpu/ethos/ethos-n78 (accessed on 29 June 2022).
75. Kneron. KL530 AI SoC. Available online: https://www.kneron.com/cn/page/soc/ (accessed on 29 June 2022).
76. Qadric. Quadric Dev Kit. Available online: https://www.quadric.io/technology/devkit (accessed on 30 June 2022).
77. Mythic. M1076 Analog Matrix Processor. Available online: https://mythic.ai/products/m1076-analog-matrix-processor/ (accessed on 30 June 2022).
78. Advanced Micro Devices, Inc. Versal AI Edge Series. Available online: https://www.xilinx.com/products/silicon-devices/acap/versal-ai-edge.html (accessed on 30 June 2022).
79. Khabbazan, B.; Mirzakuchaki, S. Design and Implementation of a Low-Power, Embedded CNN Accelerator on a Low-end FPGA. In Proceedings of the 2019 22nd Euromicro Conference on Digital System Design (DSD), Kallithea, Greece, 28–30 August 2019; pp. 647–650. [CrossRef]
80. Gondimalla, A.; Chesnut, N.; Thottethodi, M.; Vijaykumar, T.N. SparTen: A Sparse Tensor Accelerator for Convolutional Neural Networks. In Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, Columbus, OH, USA, 12–16 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 151–165. [CrossRef]
81. Im, D.; Han, D.; Choi, S.; Kang, S.; Yoo, H.J. DT-CNN: Dilated and Transposed Convolution Neural Network Accelerator for Real-Time Image Segmentation on Mobile Devices. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Washington, DC, USA, 16–20 February 2019; pp. 1–5. [CrossRef]
82. Chandramoorthy, N.; Swaminathan, K.; Cochet, M.; Paidimarri, A.; Eldridge, S.; Joshi, R.V.; Ziegler, M.M.; Buyuktosunoglu, A.; Bose, P. Resilient Low Voltage Accelerators for High Energy Efficiency. In Proceedings of the 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), Washington, DC, USA, 16–20 February 2019; pp. 147–158. [CrossRef]
83. Deng, C.; Sun, F.; Qian, X.; Lin, J.; Wang, Z.; Yuan, B. TIE: Energy-Efficient Tensor Train-Based Inference Engine for Deep Neural Network. In Proceedings of the 46th International Symposium on Computer Architecture, Phoenix, AZ, USA, 22–26 June 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 264–278. [CrossRef]
84. Jang, H.; Kim, J.; Jo, J.E.; Lee, J.; Kim, J. MnnFast: A Fast and Scalable System Architecture for Memory-Augmented Neural Networks. In Proceedings of the 46th International Symposium on Computer Architecture, Phoenix, AZ, USA, 22–26 June 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 250–263. [CrossRef]
85. Hegde, K.; Asghari-Moghaddam, H.; Pellauer, M.; Crago, N.; Jaleel, A.; Solomonik, E.; Emer, J.; Fletcher, C.W. ExTensor: An Accelerator for Sparse Tensor Algebra. In Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, Columbus, OH, USA, 12–16 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 319–333. [CrossRef]
86. Gupta, U.; Reagen, B.; Pentecost, L.; Donato, M.; Tambe, T.; Rush, A.M.; Wei, G.Y.; Brooks, D. MASR: A Modular Accelerator for Sparse RNNs. In Proceedings of the 2019 28th International Conference on Parallel Architectures and Compilation Techniques (PACT), Seattle, WA, USA, 23–26 September 2019; pp. 1–14. [CrossRef]
87. Li, Z.; Chen, Y.; Gong, L.; Liu, L.; Sylvester, D.; Blaauw, D.; Kim, H.S. An 879GOPS 243mW 80fps VGA Fully Visual CNN-SLAM Processor for Wide-Range Autonomous Exploration. In Proceedings of the 2019 IEEE International Solid- State Circuits Conference—(ISSCC), San Francisco, CA, USA, 17–21 February 2019; pp. 134–136. [CrossRef]
88. Lee, J.; Lee, J.; Han, D.; Lee, J.; Park, G.; Yoo, H.J. 7.7 LNPU: A 25.3TFLOPS/W Sparse Deep-Neural-Network Learning Processor with Fine-Grained Mixed Precision of FP8-FP16. In Proceedings of the 2019 IEEE International Solid- State Circuits Conference—(ISSCC), San Francisco, CA, USA, 17–21 February 2019; pp. 142–144. [CrossRef]
89. Huang, C.T.; Ding, Y.C.; Wang, H.C.; Weng, C.W.; Lin, K.P.; Wang, L.W.; Chen, L.D. ECNN: A Block-Based and Highly-Parallel CNN Accelerator for Edge Inference. In Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, Columbus, OH, USA, 12–16 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 182–195. [CrossRef]
90. Evans, R.D.; Liu, L.; Aamodt, T.M. JPEG-ACT: Accelerating Deep Learning via Transform-based Lossy Compression. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Virtual, 30 May–3 June 2020; pp. 860–873. [CrossRef]
91. Wu, D.; Li, J.; Yin, R.; Hsiao, H.; Kim, Y.; Miguel, J.S. UGEMM: Unary Computing Architecture for GEMM Applications. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Virtual, 30 May–3 June 2020; pp. 377–390. [CrossRef]
92. Zhao, Y.; Chen, X.; Wang, Y.; Li, C.; You, H.; Fu, Y.; Xie, Y.; Wang, Z.; Lin, Y. SmartExchange: Trading Higher-cost Memory Storage/Access for Lower-cost Computation. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Virtual, 30 May–3 June 2020; pp. 954–967. [CrossRef]
93. Srivastava, N.; Jin, H.; Liu, J.; Albonesi, D.; Zhang, Z. MatRaptor: A Sparse-Sparse Matrix Multiplication Accelerator Based on Row-Wise Product. In Proceedings of the 2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), Athens, Greece, 17–21 October 2020; pp. 766–780. [CrossRef]
94. Im, D.; Han, D.; Choi, S.; Kang, S.; Yoo, H.J. DT-CNN: An Energy-Efficient Dilated and Transposed Convolutional Neural Network Processor for Region of Interest Based Image Segmentation. *IEEE Trans. Circuits Syst. Regul. Pap.* **2020**, *67*, 3471–3483. [CrossRef]

95. Ham, T.J.; Jung, S.J.; Kim, S.; Oh, Y.H.; Park, Y.; Song, Y.; Park, J.H.; Lee, S.; Park, K.; Lee, J.W.; et al. A3: Accelerating Attention Mechanisms in Neural Networks with Approximation. In Proceedings of the 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA), San Diego, CA, USA, 22–26 February 2020; pp. 328–341. [CrossRef]
96. Park, J.; Yoon, H.; Ahn, D.; Choi, J.; Kim, J.J. OPTIMUS: OPTImized matrix MUltiplication Structure for Transformer neural network accelerator. In Proceedings of the Machine Learning and Systems, Austin, TX, USA, 2–4 March 2020; Volume 2, pp. 363–378.
97. Zhang, Z.; Wang, H.; Han, S.; Dally, W.J. SpArch: Efficient Architecture for Sparse Matrix Multiplication. In Proceedings of the 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA), San Diego, CA, USA, 22–26 February 2020; pp. 261–274. [CrossRef]
98. Weng, J.; Liu, S.; Wang, Z.; Dadu, V.; Nowatzki, T. A Hybrid Systolic-Dataflow Architecture for Inductive Matrix Algorithms. In Proceedings of the 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA), San Diego, CA, USA, 22–26 February 2020; pp. 703–716. [CrossRef]
99. Ham, T.J.; Lee, Y.; Seo, S.H.; Kim, S.; Choi, H.; Jung, S.J.; Lee, J.W. ELSA: Hardware-Software Co-design for Efficient, Lightweight Self-Attention Mechanism in Neural Networks. In Proceedings of the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 14–18 June 2021; pp. 692–705. [CrossRef]
100. Zhang, G.; Attaluri, N.; Emer, J.S.; Sanchez, D. Gamma: Leveraging Gustavson's Algorithm to Accelerate Sparse Matrix Multiplication. In Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Virtual, 19–23 April 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 687–701. [CrossRef]
101. Lu, H.; Chang, L.; Li, C.; Zhu, Z.; Lu, S.; Liu, Y.; Zhang, M. Distilling Bit-Level Sparsity Parallelism for General Purpose Deep Learning Acceleration. In Proceedings of the MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture, Virtual, 18–22 October 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 963–976. [CrossRef]
102. Jeong, G.; Qin, E.; Samajdar, A.; Hughes, C.J.; Subramoney, S.; Kim, H.; Krishna, T. RASA: Efficient Register-Aware Systolic Array Matrix Engine for CPU. In Proceedings of the 2021 58th ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 5–9 December 2021; pp. 253–258. [CrossRef]
103. Cho, B.Y.; Jung, J.; Erez, M. Accelerating Bandwidth-Bound Deep Learning Inference with Main-Memory Accelerators. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, St. Louis, MO, USA, 14–19 November 2021; Association for Computing Machinery: New York, NY, USA, 2021. [CrossRef]
104. Zeng, Y.; Sun, H.; Katto, J.; Fan, Y. Accelerating Convolutional Neural Network Inference Based on a Reconfigurable Sliced Systolic Array. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Virtual, 22–28 May 2021; pp. 1–5. [CrossRef]
105. Chen, Q.; Tian, B.; Gao, M. FINGERS: Exploiting Fine-Grained Parallelism in Graph Mining Accelerators. In Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Vancouver, Canada, 25–29 March 2023; Association for Computing Machinery: New York, NY, USA, 2022; pp. 43–55. [CrossRef]
106. Yang, T.; Li, D.; Song, Z.; Zhao, Y.; Liu, F.; Wang, Z.; He, Z.; Jiang, L. DTQAtten: Leveraging Dynamic Token-based Quantization for Efficient Attention Architecture. In Proceedings of the 2022 Design, Automation & Test in Europe Conference & Exhibition, Virtual, 14–23 March 2022; pp. 700–705. [CrossRef]
107. Wu, D.; Li, J.; Pan, Z.; Kim, Y.; Miguel, J.S. UBrain: A Unary Brain Computer Interface. In Proceedings of the 49th Annual International Symposium on Computer Architecture, New York, NY, USA, 18–22 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 468–481. [CrossRef]
108. Zadeh, A.H.; Mahmoud, M.; Abdelhadi, A.; Moshovos, A. Mokey: Enabling Narrow Fixed-Point Inference for out-of-the-Box Floating-Point Transformer Models. In Proceedings of the 49th Annual International Symposium on Computer Architecture, New York, NY, USA, 18–22 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 888–901. [CrossRef]
109. Li, Z.; Ghodrati, S.; Yazdanbakhsh, A.; Esmaeilzadeh, H.; Kang, M. Accelerating Attention through Gradient-Based Learned Runtime Pruning. In Proceedings of the 49th Annual International Symposium on Computer Architecture, New York, NY, USA, 18–22 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 902–915. [CrossRef]