Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

SAPER-AI accelerator: systolic array based power-efficient reconfigurable AI accelerator

Fahad bin MUSLIM[†‡1], Kashif INAYAT[†2], Muhammad zain SIDDIQI[†1], Safiullah KHAN[3], Tayyeb MAHMOOD[4], Ihtesham ul ISLAM[5]

[1]*Faculty of Computer Science and Engineering, GIK Institute 23460, Pakistan*

[2]*Barcelona Supercomputing Center, Spain*

[3]*Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BX, United Kingdom*

[4]*Nextwave, Korea*

[5]*National University of Sciences and Technology, Pakistan*

[†]E-mail: fahad.muslim@giki.edu.pk; kashif.inayat@bsc.es; zain.siddiqi@giki.edu.pk

Received Sep. 21, 2024; Revision accepted Jan. 24, 2025; Crosschecked

Abstract: Deep learning (DL) accelerators are critical for handling the growing computational demands of modern neural networks. Systolic array (SA) based accelerators consist of a 2D mesh of processing elements (PE) working cooperatively to accelerate matrix multiplication, a fundamental operation in DL. The power efficiency of such accelerators is of primary importance especially considering the edge AI regime. This work presents the SAPER-AI accelerator, an SA accelerator with power intent specified via a unified power format representation in a simplified manner with negligible micro-architectural optimization effort. Our proposed accelerator switches off rows and columns of PEs in a coarse-grained manner, thus leading to SA micro-architecture complying with the varying computational requirements of modern DL workloads. Our analysis demonstrates enhanced power efficiency ranging between 11% and 25% for the best case 32×32 and 64×64 SA designs, respectively. Additionally, the power delay product (PDP) exhibited a progressive improvement of around 6% for larger SA sizes. Moreover, a performance comparison between the MobileNet and ResNet50 models indicated generally better SA performance for the ResNet50 workload. This is due to the more regular convolutions portrayed by ResNet50 that are more favored by SAs, with the performance gap widening as the SA size increases.

Key words: AI accelerators; ASIC design; Systolic arrays; Low power designs
https://doi.org/10.1631/FITEE.2400867 **CLC number:** TP

1 Introduction

The meteoric rise in deep learning (DL)-based solutions has revolutionized a variety of domains such as image processing, pattern recognition and transportation. To make full use of the benefits that such DL models can accord, huge operational costs need to be incurred due to the computational complexities accompanying such models (Yüzügüler et al., 2023). Therefore, specialized DL accelerators are necessary to offer enhanced computational prowess while still being energy efficient. Several such accelerators have been proposed on both sides of the DL continuum, i.e., the cloud (Bobda et al., 2022; Li et al., 2023) and the edge (Seshadri et al., 2022; Loh et al., 2024). Deep neural networks (DNNs) incorporate matrix multiplication as the primary primitive, which in turn, fortunately offers a lot of parallelism, and its acceleration is imperative to achieve the tremendous processing demands of the DL workloads (Muslim et al., 2024).

Several general matrix multiplication (GEMM) accelerators found in the literature are based on sys-

‡ Corresponding author

tolic arrays (SAs) (Jouppi et al., 2017; Song et al., 2019; Lai and Zhang, 2024). SA is a two-dimensional mesh of processing elements (PEs) with each PE being fed inputs from the left and top sides. Each PE performs a multiplication and accumulation (MAC) operation every clock cycle, and the partial result and the input are fed to the neighboring PE via pipeline registers. In this way, the PEs collaborate to offer enhanced parallelism in processing data efficiently and accelerate the DNN computation. A typical SA architecture is depicted in Fig 1. Multiplier `X` is indicated on the left while the multiplicand `Y` and the bias term `W` on the top side of SA in Fig. 1

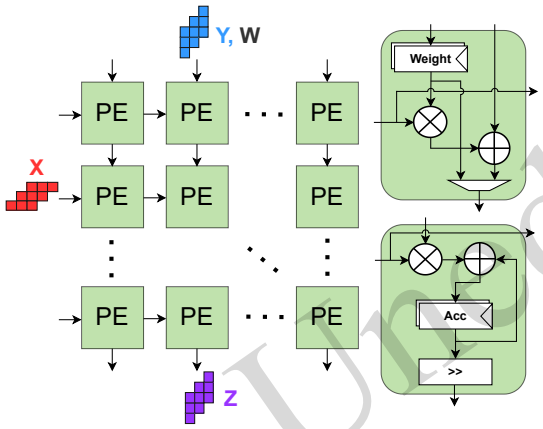

Fig. 1 A typical SA architecture with output and weight stationary data flows. PE, processing element; SA, systolic array.

One of the main factors necessitating the usage of customized architectures for DL acceleration is the resulting improvement in the energy efficiency offered by the accelerators. Such customized hardware must comply with the tight power budget, which is important both at the cloud and the edge. However, the power efficiency of such designs at the edge due to the constrained power availability is of even greater importance (Kim et al., 2020). Thus, a lot of research has been done in recent times to improve the power efficiency of such hardware accelerators. The research found in the literature mainly targets complex micro-architectural optimizations, e.g., by reducing expensive memory access optimizations or by quantization of the networks leading to approximate computing, thus leading to enhanced energy savings (Chen et al., 2016; Moons et al., 2016).

Moreover, the modern DNNs are accompanied by increased sparsity, much like their biological counterparts and can generalize well also when compared to their denser versions (Hoefler et al., 2021; Guo et al., 2024). Owing to their reduced memory footprints, enhanced power efficiency and lightweight implementation more suited to the inference regime, such networks are primarily suited for edge AI devices (Hoefler et al., 2021). The sparsity, however, is counter productive when the DNN implementation via SA-based accelerators is concerned. This is due to the fact that the SAs are well suited for dense, regular computation patterns accompanied by predictable dataflow in the array (Xu et al., 2021a). The irregularity in data accesses as well as in computation patterns causes several PEs to remain idle, thus consuming power while not doing any useful computation, thus adversely affecting the power efficiency (Chen et al., 2016; Xu et al., 2023). The impact of various types of convolutions on the utilization of SAs is depicted graphically in Fig. 2. Figure 2a shows the typical GEMM, while the matrix-vector multiplication (mimicking depthwise convolution) is depicted in Fig. 2b. Finally the impact on SA utilization when the SA size is large, considering depthwise convolution, is shown in Fig. 2c. The SA utilization is indicated by the ratio of shaded PEs to total PEs in Fig. 2. The interdependence between the SA size and the type of convolution operation with respect to SA utilization can clearly be seen in the figure as well.

The aforementioned discussion leads us to the primary motivation behind this research effort. This includes the ability to reconfigure SAs meant for dense computations in accordance with the required computational complexity of various network layers and power gate the idle PEs in bulk (coarse-grained manner). This shall ensure enhanced power efficiency at the cost of additional logic caused by the introduction of the power gating logic. The architectural details of two neural network architectures, i.e., MobileNet and ResNet50 (both trained using the ImageNet dataset), considered for design evaluation in this work are given in Table 1. We observe a reduction in the output feature map (OFMap) size as the number of network layers increases (Xu et al., 2021a). The depthwise and pointwise convolutions in the MobileNet layers can also be observed. The frequent transition between these two convolutions has a profound impact on the SA performance eval-

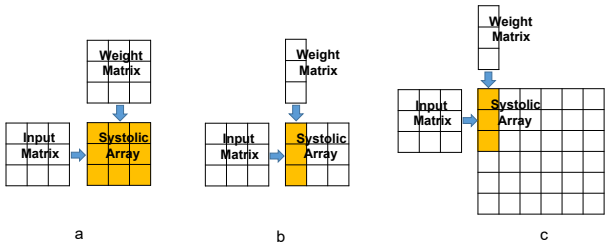

Fig. 2 (a) Typical GMEM, (b) Matrix-vector Multiplication, and (c) Matrix-vector multiplication with large SA size. SA, Systolic Array.

uation considering MobileNet, as will be presented in Section 4.

Power gating is achieved by writing the power intent for the SA accelerator in an IEEE standard unified power format (UPF). The power management logic is inserted at a coarse-grain level and adjusted in accordance with the layerwise computational demands of the considered DNN models. Power intent description via UPF allows power logic to be handled separately from the logic intent description (Chadha and Bhasker, 2012). This, in turn, leads to simplified power inference with trivial design changes required to be made at the Register Transfer Level (RTL), mainly to provide power control signals to control the insertion of low power logic after logic synthesis. In comparison to our previous work (Inayat et al., 2023) targeting fine-grained power gating by shutting down the idle multipliers of the MAC units inside the PEs, this work targets a coarse-grained approach. This approach better fits the sparse nature of modern DNN workloads. Moreover, that effort included microarchitectural optimizations to divide the multipliers in the MAC units into smaller submultipliers to increase the power gating instances (at the cost of increased area). In comparison, this work purely exploits the network architecture without having to rely on any expensive microarchitectural modifications, apart from the requirement to write the power management module at the RTL and instantiate it inside the top SA module.

The major contributions targeted in this work are as follows:

- ASIC implementation of a 32×32 and 64×64 reconfigurable SA-based AI accelerator termed as SAPER-AI accelerator, to cater to the varying workloads of the considered DNN model layers. This ensures the full utilization of the active PEs of the underlying SA designs at any given time.
- Coarse-grained power gating of row and column PEs of the SA based on the varying computational requirements of various network layers of the DNN workload. The main aim of the power optimization strategy is to achieve power savings with no expensive micro-architectural modifications.
- Realistic power analysis of the proposed design cases by utilizing actual workloads of some DNN architectures, i.e., MobileNet and ResNet50.
- A detailed comparison of the proposed SA design based on power, performance, area (PPA), and power delay product (PDP) parameters.

The manuscript is organized in the following manner. Section 2 presents some relevant literature survey on the power efficiency of AI accelerators. The detailed system model, including the micro-architectural details, the design choices and the power analysis framework details, is presented in Section 3. Finally, the results are presented in Section 4, while Section 5 presents the concise conclusions that can be drawn from this research effort. Section 6 discusses some prospective future directions in which this work can be extended.

2 Related Work

This section presents a mix of some recent works targeting power-efficient SA-based accelerators, while also specifically dealing with reconfigurable accelerators that can adjust their features to cater to evolving DNN workloads.

(Moghaddasi and Nam, 2024) presented a serial/parallel and Octet serial/parallel SA (SPSA and OSPSA) approach. The designs presented improve the DNN computational efficiency by activating serial processing in an adjustable manner with respect to the runtime and layerwise precision. This is done in accordance with the computation required by the DNN model. While both the activations and weights fed to the PEs are done in a parallel manner in traditional SAs, their proposed approach introduces support for serializing the activation matrices while keeping the weight matrices input in a parallel manner. The OSPSA design further improves the throughput by processing a column of eight simultaneous bits of different activations in the same bit po-

Table 1 Structure of MobileNet and ResNet50 models.

Architecture	Layer name	Kernel size // OFMaps Size
MobileNet	Conv1	3x3 // 112x112
	Conv2_dw/pw	3x3 / 1x1 // 112x112
	Conv3_dw/pw	3x3 / 1x1 // 56x56
	Conv4_dw/pw	3x3 / 1x1 // 56x56
	Conv5_dw/pw	3x3 / 1x1 // 28x28
	Conv6_dw/pw	3x3 / 1x1 // 28x28
	Conv7_dw/pw	3x3 / 1x1 // 14x14
	Conv8_dw/pw ×5	3x3 / 1x1 // 14x14
	Conv9_dw/pw	3x3 / 1x1 // 7x7
ResNet50	Conv1	7x7 // 112x112
	Conv2_3	1x1 / 3x3 / 1x1 // 56x56
	Conv3_4	1x1 / 3x3 / 1x1 // 28x28
	Conv4_6	1x1 / 3x3 / 1x1 // 14x14
	Conv5_3	1x1 / 3x3 / 1x1 // 7x7

OFMap, output feature map.

sition. Energy consumption is reduced via the simpler serial/parallel PE architecture and the replacement of complex multipliers with simpler and cost-effective serial circuits. Other bit-serial approaches are also presented in (Lee et al., 2018; Ryu et al., 2019), which obviously offer advantages in terms of interconnections, area, and power. However, this gain traditionally comes with a sacrifice in throughput of the underlying SA-based designs.

The authors in (Inayat et al., 2023) proposed a power intent systolic array (PI-SA), wherein they rely on UPF-based power intent, similar to our present work. However, power gating was done in a more fine-grained manner by targeting multipliers for power shutoff in case of inactivity. This is different from what we are doing, i.e., using coarse-grained power gating. Power efficiency is further improved by splitting the multipliers into smaller parallel submultipliers, thus increasing the power shutoff instances. Moreover, the power analysis in that work includes introducing zero entries randomly and not necessarily indicating the data traffic of actual DNN workloads considered in this work.

The authors in (Xu et al., 2021b) presented the concept of a heterogeneous systolic array architecture (HeSA) to resolve the issue of suboptimal SA performance when processing compact convolutional neural network (CNN) models, such as depthwise convolution. This work introduces heterogeneous PEs supporting multiple dataflows to cater to changing DNN workloads. Since the structural changes are happening to the PEs within the SAs, the traditional SA structure remains intact. While dealing with standard convolution, the HeSA acts as a traditional SA, but for the depthwise convolution, the SA adjusts the dataflow by utilizing the heterogeneous PEs, thus resulting in significant energy savings when compared to naive SA architecture. While this work proposes a way around the depthwise convolutions which are also considered in this work, the approach is considerably different and arguably more complicated than our approach.

Another approach to achieve energy efficiency in AI accelerators is via approximate computing based on the quantization of weights and inputs at each layer, as presented in (Moons et al., 2016). Other approximate computing approaches based on dynamic voltage and frequency scaling (DVFS) to vary the threshold voltage leading to supply voltage variation for differing precision requirements are presented in (Moons et al., 2017a,b). These works, however, achieve energy efficiency at the cost of loss in precision, which may be crucial in some DNN applications, such as medical imaging and diagnostics, and military and defense.

Thus, the presented literature can be summarized by inferring that energy efficiency in DL accelerators is of prime importance. This is usually achieved by complicated microarchitectural modifications and other approximation approaches, which, in turn, lead to a loss of accuracy. Based on the discussion above, the presented literature can be broadly summarized in terms of "optimization strategies,", "accuracy loss," and "optimization effort". Such a comparison of the presented literature with the proposed work in these aspects is given in Table 2. This work intends to improve the accelerator energy efficiency by specifying power intent via UPF with trivial changes in the SA structure. The optimizations can be adjusted according to the ac-

tual DNN workload, and significant energy savings can be achieved as demonstrated by performance analysis considering a couple of widely used DNN models.

3 System Model

This section details the system architecture, indicating primarily how the power intent description process will vary the SA architecture. We then present the 32×32 and 64×64 SA accelerator designs considered in this work. Each design has adjustable PE configurations to cater to the varying DNN workloads. Finally, we present the details of our power analysis framework, considering MobileNet and ResNet50 DNN workload models.

3.1 Bird's Eye Illustration of the System

A high-level abstraction of the architecture of the 64×64 SA top module with various module instantiations is shown in Fig. 3. The power management unit (PMU) is an RTL module basically responsible for providing the control signals to shut off the power domains. These power domains are labeled as `PD_X`, where `X` ranges from 2 to 5, indicating the switchable power domains. Power domains are basically a collection of PEs that operate at the same operating conditions, i.e., all the power domains collectively constitute a power-aware SA design. Another variable of interest is the `mesh_select` signal, which (as the name implies) decides which power domains (or a collection of PEs) are being put to sleep at any instant. When the `mesh_select` variable is set to 0, all the switchable domains are put to rest, thus allowing only the 7×7 SA size to compute actively, i.e., the so-called default power domain (`PD_default`). The term "default" indicates that this power domain will never be switched off under any circumstances. The PMU module is also part of the default power domain, as it is responsible for providing the necessary power control signals and must be active at all times. Similarly, with `mesh_select` equaling 1, the 14×14 sized SA computes actively while the rest of the PEs are put to sleep. The rest of the active SA sizes selected via the `mesh_select` variable are 28×28, 56×56, and the fully loaded 64×64 for `mesh_select` values of 2, 3, and beyond, respectively. We do not depict the 32×32 SA since it has a similar architecture. The `mesh_select` values of 0, 1, and 2 select between 7×7, 14×14, and 28×28 active SA sizes, respectively, with the full blown active SA size being 32×32. The PMU is meant to provide the power control signals in an appropriate manner. We do not discuss its detailed logic but inquisitive readers are encouraged to explore the same in other works reported in (Qamar et al., 2017) and (Inayat et al., 2023).

At the gate level, committing power intent results in the insertion of low-power logic in the design (Qamar et al., 2017). This logic includes low-power cells, such as isolation cells, to isolate the default power domain from the floating output values of the power-gated domain. These cells are usually placed at the intersection of the two domains (at the output of the switchable domain). Another class of low-power cells is the retention cells. These are used to save the state of some sequential logic in the power-gated domain before it is put to sleep. These cells are inserted if the logic synthesis tool senses the need for them. Furthermore, power switch header/footer cells are used to cut off the supply to the switchable domain and are usually placed in the design after physical implementation (Chadha and Bhasker, 2012). A generalized view of a typical power-aware hardware with a single default Always-ON domain and a power shut-off (PSO) domain similar to the one created in this work is shown in Fig. 4. The PMU module indicating the power control signals, `iso` (isolation) and `ret` (state retention), and the resulting low-power cells i.e., `ISO` and `RET`, can be clearly seen as well. Both the Always-ON and the PSO domains contain group of PEs in our case.

3.2 Detailed Microarchitecture of the Designs

This section presents the detailed microarchitectural representation of the power-aware 32×32 and 64×64 SAs. The 32×32 design is depicted in Fig. 5, while its 64×64 counterpart is shown in Fig. 6. A multiplexer at the input of the SA in both cases is used to provide the inputs to the active PEs. The SA inputs as well as the PMU signals in both figures are color-coded in accordance with the active power domains at any specific time. The low-power logic depicted earlier in Fig. 4 with respect to various power domains is inserted by the logic synthesis tool. This logic ensures that the appropriate power domains are active at any given time as dictated by the DNN computation workload. Sequential logic is

Table 2 Broader Comparative Analysis with Relevant Literature

Design	Optimization strategy	Accuracy loss	Effort
(Lee et al., 2018; Ryu et al., 2019) (Xu et al., 2021b; Moghaddasi and Nam, 2024)	Microarchitectural optimizations	No	High
(Inayat et al., 2023)	PG (Fine) & Microarchitetural optimizations	No	Medium
(Moons et al., 2016)	Approximate computing (Quantization)	Yes	Medium
(Moons et al., 2017a,b)	Approximate computing (DVFS)	Yes	Low
This work	Coarse-grained power gating	No	Low

DVFS, dynamic voltage and frequency scaling.

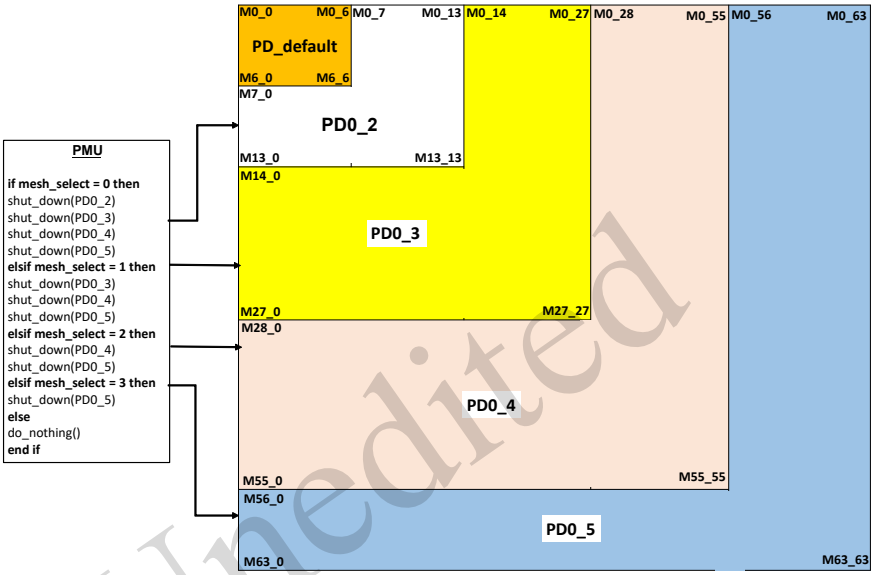

Fig. 3 Abstracted Illustration of Power-Aware 64×64 SA Design, SA, systolic array.

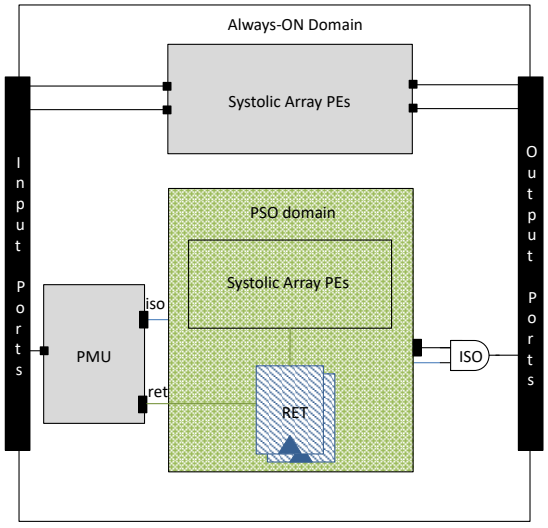

Fig. 4 Generalized Power-Aware Hardware Design. PEs, processing elements; PMU, power management unit; PSO, power shut-off.

also depicted at the output of the SA in Figs. 5 and 6. This is necessary to add the appropriate delay in each of the low-power SA operations to ensure that the SA output is always fed to any succeeding components in an amicable fashion. The only exception, as far as the registered SA outputs are considered, is the full-blown SA computation that would always be active had the power logic not been used in the design. Moreover, it should also be noted that even though a single register is shown to store the various active PE outputs, it is only to simplify the depiction. In reality, the number of registers differs depending on the active SA size. As an instance, for the 32×32 case depicted in Fig. 5, with 7×7 active PEs, we use 50 registers, while in the case of 64×64 SA depicted in Fig. 6 and considering 7×7 active PEs, we use 114 registers. To ensure appropriate SA outputs at all times, the general number of registers required to store the active SA outputs, considering the active SA size of $\mathtt{n} \times \mathtt{n}$ and the full-blown SA size of $\mathtt{N} \times \mathtt{N}$, is $2(\mathtt{N} - \mathtt{n})$. The register count required at the SA output for different cases is given in Table 3.

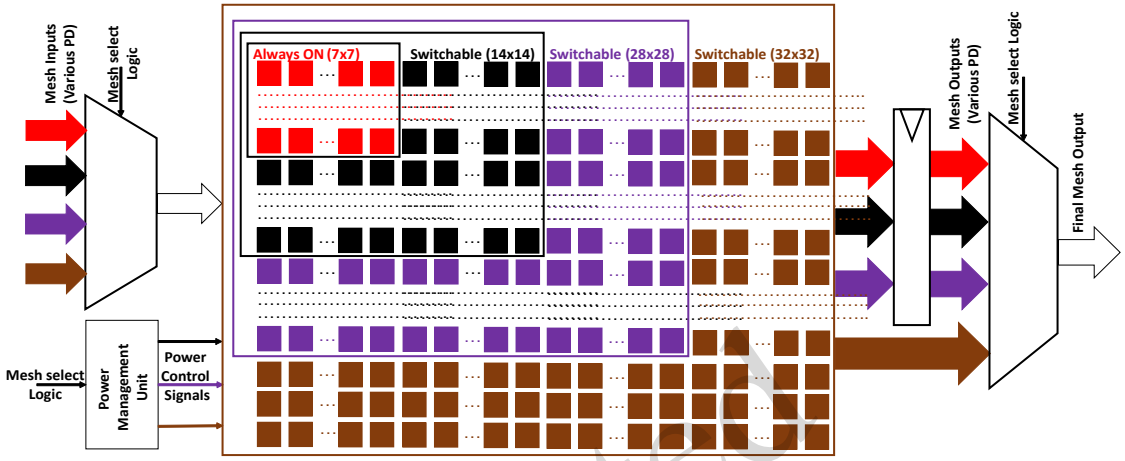

Fig. 5 Detailed View of 32×32 SA with Power Control Signals. SA, systolic array.

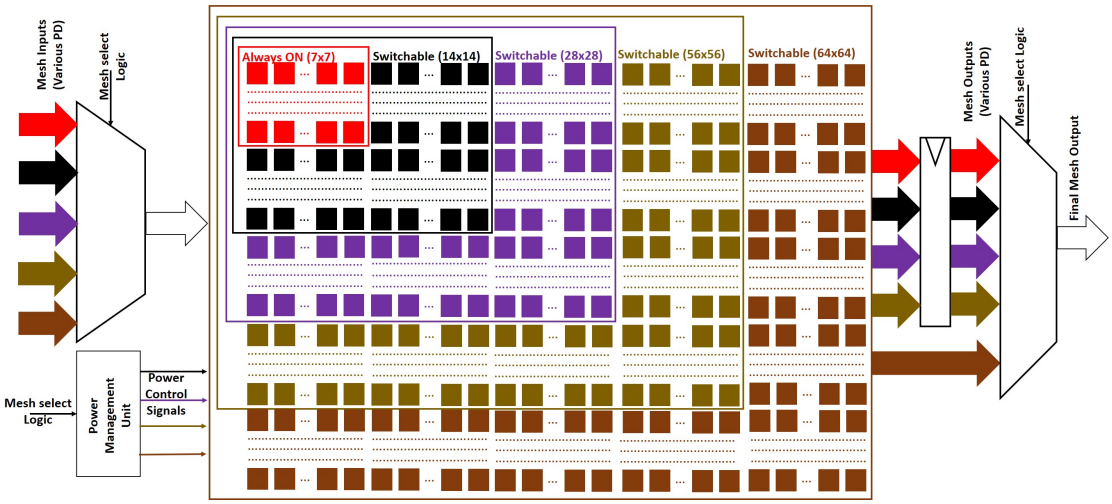

Fig. 6 Detailed View of 64×64 SA with Power Control Signals. SA, systolic array.

Table 3 Number of Registers required at SA output for Various sizes

Total SA size ($N \times N$)	Active SA size ($n \times n$)	# O/P registers ($2[N-n]$)
32×32	7×7	50
	14×14	36
	28×28	8
64×64	7×7	114
	14×14	100
	28×28	72
	56×56	16

SA, systolic array.

Table 4 Activity profile of various feature sizes of DNN workload

Model	Output size	Activity (%)
MobileNet (1278M cycles)	56×56	15.63
	28×28	19.5
	14×14	20.78
	7×7	11.16
ResNet50 (109M cycles)	56×56	14.56
	28×28	25.13
	14×14	21.02
	7×7	13.12

DNNs, deep neural network.

3.3 Power Analysis Framework

This section details the power evaluation framework used in this work. The power-aware SA designs proposed in this work are based on the baseline design generated by the Gemmini SA generator, as presented in (Genc et al., 2021). The generated SAs are evaluated by running DNNs such as MobileNet and ResNet50 as per the baseline design evaluation framework presented in (Genc et al., 2021). We performed initial simulations in the baremetal environment using our recently developed in-house FPGA-accelerated simulation environment called the "Quickloop" (Inayat et al., 2024). These simulations were meant to extract the number of clock cycles taken by each layer of the DNN workload, i.e., to do activity profiling of the DNN workloads. This number was then used to estimate the index numbers (with respect to the whole DNN activity index) to perform power analysis while complying with the actual workloads of the DNN models considered in this work.

Furthermore, the gate-level netlist produced after logic synthesis was also simulated to generate the switching activity information of various signals. This information was captured using the switching activity interchange format (SAIF) file that was then annotated to the gate-level netlist to perform the power analysis. The activity profile for various computation sizes of both the considered DNN workloads as obtained from FPGA-accelerated simulation is provided in Table 4. The total number of clock cycles taken by the FPGA-assisted SA simulation of both networks is also indicated in the table. The MobileNet simulation seems to be taking more clock cycles to complete. This may be attributed to the frequent switching between depthwise and pointwise convolutions which in turn causes dataflow patterns to change significantly, thereby leading to under utilization of the SA. Moreover, the delay caused by excessive memory accesses and synchronization overheads due to frequent switching of convolutions patterns in the MobileNet model also adversely impacts the simulation cycle count (Lym and Erez, 2020; Xu et al., 2021a).

4 Results

This section discusses the details regarding the evaluation environment used in this work. This is followed by a detailed performance analysis of the developed 32×32 and 64×64 SA designs considering, the DL workloads.

4.1 Evaluation Environment

All the designs have been synthesized using the SAED `32nm` technology library supporting low-power designs. We have employed various Synopsys tools, i.e., Design Compiler for synthesis, VCS for simulation and Primetime for power analysis. Functional simulation has been performed both at the RTL and the gate-level design representation. The power measurement has been done at `3ns` clock period. All the analysis has been done on a Linux machine with `128GB` of memory. This memory constraint imposes a limit on the largest SA size being deployed, especially considering the power-gated designs. This is the reason behind the maximum SA size of 64×64 considered in this work.

4.2 32×32 SA Performance Analysis

The performance analysis for the 32×32 designs in terms of delay and area, with and without power intent, is given in Table 5. The cases with the best delay and area numbers are emboldened in Table 5. The insertion of low-power cells has an adverse effect on the delay and area of the design. The main culprits, as far as the area increase of the power-gated

Table 5 32×32 SA performance analysis wrt DNN models

Performance parameters	MobileNet		ResNet50	
	No UPF	With UPF	No UPF	With UPF
Delay (ns)	**1.56**	1.91	**1.56**	1.91
Area (mm^2)	**7.74**	8.41	**7.74**	8.41
Power (W)	1.094	**0.98**	1.077	**0.95**
PDP (W.ns)	**1.71**	1.87	**1.68**	1.81

DNN, deep neural network; PDP, power delay product; SA, systolic array; UPF, united power format.

designs is concerned, are the sequential area and the interconnect area. This is because the power gating results in excessive sequential logic in the form of output registers, in addition to an increase in resulting interconnections. These contribute significantly to the delay of the power-aware designs as well as the delay caused by the additional low-power cells and the interconnects that the insertion of these cells shall result in.

As far as the power analysis is concerned, the 32×32 SA design shows power saving of around 11% in the case of the MobileNet and around 12% in the case of the ResNet50 workload, as depicted in Table 5. The combined impact of both the power and delay can be captured in terms of the PDP. In terms of the PDP, our power-gated design worsens by around 9% in the case of the MobileNet workload and by around 8% in the case of the ResNet50 workload. Thus, besides improving the power consumption by around 11%-12% for realistic DNN workloads, our developed power-aware 32×32 SA design does not perform all that well in the other performance parameters. This prompted us to explore bigger-sized SAs, as 64×64 and 128×128 SAs are not that uncommon, especially for processing large-scale matrix multiplications that are fundamental to DNNs (Chen et al., 2014, 2016; Gao et al., 2017; Jouppi et al., 2017; Chen et al., 2019). Due to memory limitations of our evaluation setup, we could not synthesize the 128×128 SA design and were hence restricted to the 64×64 SA design, that is presented next.

4.3 64×64 SA performance analysis

The performance evaluation for the power-aware 64×64 SA accelerator is presented in Table 6. The best-case performance parameters in each case are emboldened, similar to Table 5. The trend is the same as that of the 32×32 case when the non-power-aware and power-aware cases are compared, i.e., the delay and area increases with a decrease in the power consumption. The delay cost in the case of power-gated 64×64 SA is around 25%, while the associated area cost is around 11%. The power-gated 64×64 design shows slightly higher delay compared to the power-gated 32×32 design. The delay of the 32×32 and 64×64 SA designs, considering the same power awareness regimes, is almost the same, however. This is because of the critical path in both the 32×32 and 64×64 cases being constructed within a single PE, which is technically the same. This is because this work does not change the PE micro-architecture but instead makes slight micro-architectural changes to the SA itself. There is a 4× increase in the area of the 64×64 design as compared to the 32×32 design, however, due to a massive increase in the computational and interconnect logic in the case of the 64×64 design.

The delay in the case of 64×64 SA worsens by around 25%, while the area by around 11% as a result of the lower power logic. The power, however, improves by 22% for the MobileNet case and by about 25% for the ResNet50 case. More significantly, the PDP shows an improvement of around 2% for the MobileNet case and around 6% for the ResNet50 case when considering the power-aware 64×64 SA design. This positively cements our hypothesis that the proposed optimizations are more useful when considering larger SA designs. Such larger arrays are beneficial for different applications such as high-performance cloud-based AI accelerators and inference of large models, e.g., BERT and GPT.

4.4 DNN model comparative performance

This section performs a comparative assessment (across both the power-aware and non power-aware) of the best case performance of the 32×32 and 64×64 SA designs. The best case performance for the 32×32 SA design is depicted graphically in Fig. 7. The area and delay performance is dependent on the design itself and does not vary with the DNN workloads and hence is the same for both MobileNet and ResNet50 workloads. They are still, however, included in the

Table 6 64×64 SA performance analysis wrt DNN models

Performance Parameters	MobileNet		ResNet50	
	No UPF	With UPF	No UPF	With UPF
Delay (ns)	**1.56**	1.96	**1.56**	1.96
Area (mm^2)	**30.8**	34.2	**30.8**	34.2
Power (W)	3.72	**2.9**	3.6	**2.7**
PDP (W.ns)	5.80	**5.68**	5.62	**5.29**

DNN, deep neural network; PDP, power delay product; SA, systolic array; UPF, united power format.

analysis for the sake of completeness. The best-case power results are obtained for the 32×32 case by the power-aware design, while the non-power-aware case achieves the best PDP performance. The two DNN models can not be separated much in this case, with slightly better performance achieved by ResNet50 as far as the design power consumption is concerned.

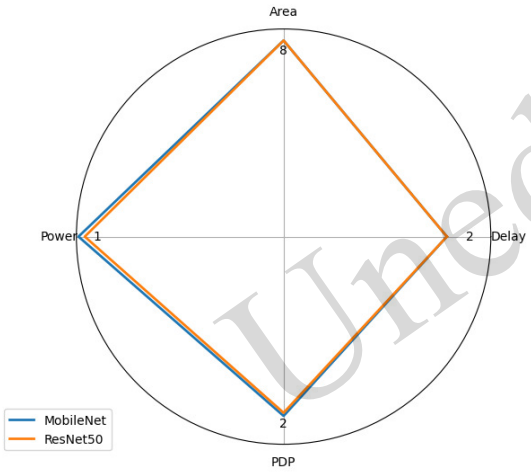

Fig. 7 Best-case performance analysis of the 32×32 SA design. PDP, power delay product; SA, systolic array.

The best-case performance results for both the DNN models in the case of the 64×64 SA design are depicted in Fig. 8. Just like the 32×32 SA design, the 64×64 array also shows better performance in terms of power and PDP for the ResNet50 model. The model-wise difference in performance, however, is much more prominent as compared to the 32×32 array. This is due to the consistent switching between pointwise and depthwise convolutions portrayed by MobileNet that the SA architecture is not very good at handling (Park et al., 2024). This structural difference between the DNN models seems to cause a larger variation in performance as the SA size goes up. Moreover, the best-case power and PDP performance in the case of 64×64 SA, considering both the DNN models for the power-aware case, is much better as compared to the non-power-aware case. This indicates that the proposed design subsequently performs much better in terms of DL acceleration considering larger SA sizes.

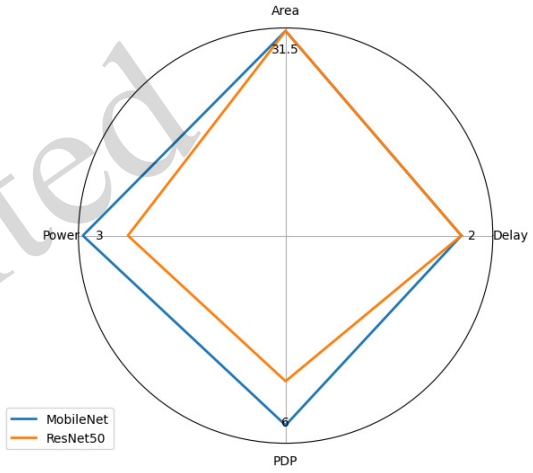

Fig. 8 Best-case performance analysis of the 64×64 SA design. PDP, power delay product; SA, systolic array.

4.5 State-of-the-art Comparison

A conceptual comparison of the proposed work with similar state-of-the-art endeavors is presented in Table 2. Those works, while being relevant, are still considerably different in approaches, and thus, a fair numerical comparison with all the works is not possible. We have thus chosen a subset of those works based on the same baseline naive SA design presented in (Genc et al., 2021) for a numerical comparison with our proposed work. Furthermore, since our proposed work does not utilize approximation of any sort, we include only the literature exhibiting no loss of accuracy. The analysis presented here must be considered in combination with other factors such as "optimization effort," and "real-time DL workload" consideration to assess where the proposed work would fit in the current design landscape. The comparison is tabulated in Table 7, which shows

Table 7 Numerical comparison with the state-of-the-art work

Design	*ΔArea*	*ΔEnergy*	Max clk	Max SA size	DNN models	Tech
(Xu et al., 2021b)	None	10%	500MHz	32×32	MobileNet, MixNet	45nm
(Inayat et al., 2023)	32%	56%	250MHz	64×64	None	32nm
This work	↑ 10%	6%	333MHz	64×64	ResNet50, MobileNet	32nm

DNN, deep neural network; SA, systolic array.

that the "maximum clock frequency," the "SA size," and the "technology size" considered in this work are all in line with the parameters considered in similar studies.

Table 7 primarily indicates the change in area ($\Delta area$) and the change in energy ($\Delta energy$), the proposed optimizations in the cited works cause as compared to the naive SA representation. While (Inayat et al., 2023) portrayed better performance as far as the area and the energy consumption are concerned, it uses random zero inputs to cause fine-grained powering off of the PEs. The values thus do not represent the real workload handled by modern DL frameworks. In comparison, the work presented by (Xu et al., 2021b) considered real DL workloads, with energy savings almost equal to those offered by our proposed design. Our work, however, causes a slight degradation of area due to the additional low-power logic and the sequential logic at the design output. The numbers presented in Table 7, when considered in combination with the optimization effort requirement of each approach, emphasize the significance of our proposed work compared to the relevant state-of-the-art research.

4.6 Limitations

While we have analyzed the performance of our designed accelerator on MobileNet and ResNet50 models, we realize the enhanced dominance of transformer-based models such as GPT, in the prevailing AI workloads. Transformer models rely heavily on matrix multiplications in self-attention and feedforward networks, but the matrices involved are considerably large and not as structured as in convolutions (Chang and Kim, 2024). Thus, scaling the proposed SA design to transformer-based models would pose significant challenges in dealing with memory access bandwidth and dynamic dataflow requirements of these models (Amirshahi et al., 2023, 2024).

Besides, we have a constraint on the maximum size of the SA, i.e., 64×64, in this work. The large transformer models may involve splitting the computations into multiple blocks (with the maximum SA size limitation), leading to idle PEs due to partial computations. The power (especially dynamic power) may therefore increase because of the resulting frequent switching. This mismatch in workload granularity may require fine-grained power gating at the PE level to achieve better energy efficiency but at the cost of higher power logic. Additionally, the performance evaluation in this work, as well as other similar works presented earlier, was based on the Gemmini framework proposed in (Genc et al., 2021). This framework utilized MobileNet and ResNet50 models for evaluating performance, and thus, we used the same workloads for a fair comparison to other state-of-the-art models.

To summarize, extending the proposed work to incorporate the transformer models would require several low-level interventions, such as on-chip caching and tiling strategies to reduce memory access overheads as well as compression mechanisms to reduce attention weights and activations. This would in turn add additional optimization effort, which is not what the main scope of this activity was, i.e., to introduce simplified power intent in the SA-based accelerator and analyze its performance on DL workloads. We thus intend to address these significant challenges in our future endeavors to enhance the scalability of our proposed work towards transformer models for targeting applications such as natural language processing and computer vision.

5 Conclusion

In this research endeavor, we propose a power-aware SA-based AI accelerator termed as the SAPER-AI accelerator. This accelerator is meant to exploit the structural changes in modern DNN workloads and reconfigurably select active PEs to cater to the changing workloads. Unlike similar efforts found in the literature that rely on complex micro-

architectural optimizations or fine-grained power optimizations, the proposed accelerator relies on much simpler UPF-based power intent descriptions to put unutilized PEs to sleep in a coarse-grained manner. The accelerator involves modifying the naive Chisel-generated SA trivially by introducing a power management module that is responsible for providing the appropriate power control signals. The low-power cells introduced in the design after logic synthesis ensure the power intent integrity of the design. The same is validated by the post-synthesis power-aware simulation as well.

Evaluating the performance of the proposed designs with scaling demonstrates that the optimizations generally cause area and delay penalties due to the additional low-power logic. Power and PDP evaluation, considering MobileNet and ResNet50 DL workloads, indicate that both the 32×32 and 64×64 SA designs are more power-efficient comparatively in the power-aware case. The efficiency, however, improves further as the SA sizes increase. Moreover, the PDP of the power-aware 64×64 SA is also considerably better when compared with the non-power-aware case, thus demonstrating the effectiveness of the proposed optimizations, especially for larger SAs. Finally, the proposed accelerator provides better performance on the ResNet50 workload compared to the MobileNet case for larger SAs owing to greater comparative uniformity in convolutions of ResNet50 that are more favored by the underlying SA architecture.

6 Future Work

As the future scope of work, further research is required to validate the enhancement in the performance of the proposed optimizations with increasing SA sizes. This could not be obtained in this work due to memory capacity limitations in our evaluation setup. Moreover, other more complex power optimization techniques based on UPF, such as dynamic voltage scaling, can also be used to optimize the design power even more. We can also explore some simpler microarchitectural optimizations to keep area and delay values in check while optimizing power. Additionally, we intend to scale the proposed work towards a power optimized SA-based accelerator design targeting transformer-based models. This will require the necessary micro-architectural interventions to circumvent the challenges posed by the complicated matrix operations constituting the transformer models as well as by the considerably larger number of parameters of such models.

Contributors

Fahad Bin MUSLIM conceptualized the research. Fahad Bin MUSLIM and Kashif INAYAT implemented the idea and did the analysis. Fahad Bin MUSLIM and Muhammad zain SIDDIQI drafted the paper. Safiullah KHAN and Tayyeb MAHMOOD helped organize the paper while Ihtesham ul ISLAM revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

References

Amirshahi A, Klein JAH, Ansaloni G, et al., 2023. Ticsat: Tightly-coupled systolic accelerator for transformers. Proceedings of the 28th Asia and South Pacific Design Automation Conference, p.657-663.

Amirshahi A, Ansaloni G, Atienza D, 2024. Accelerator-driven data arrangement to minimize transformers runtime on multi-core architectures. 15th Workshop on Parallel Programming and Run-Time Management Techniques for Many-Core Architectures 13th Workshop on Design Tools.

Bobda C, Mbongue JM, Chow P, et al., 2022. The future of fpga acceleration in datacenters and the cloud. *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, 15(3):1-42.

Chadha R, Bhasker J, 2012. An ASIC low power primer: analysis, techniques and specification. Springer Science & Business Media.

Chang SW, Kim DS, 2024. Scalable transformer accelerator with variable systolic array for multiple models in voice assistant applications. *Electronics*, 13(23):4683.

Chen YH, Emer J, Sze V, 2016. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. *ACM SIGARCH computer architecture news*, 44(3):367-379.

Chen YH, Yang TJ, Emer J, et al., 2019. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(2):292-308.

Chen Y, Luo T, Liu S, et al., 2014. Dadiannao: A machine-learning supercomputer. 2014 47th Annual IEEE/ACM International Symposium on Microarchitecture, p.609-622.

Chen Y, Chen T, Xu Z, et al., 2016. Diannao family: energy-efficient hardware accelerators for machine learning. *Communications of the ACM*, 59(11):105-112.

Gao M, Pu J, Yang X, et al., 2017. Tetris: Scalable and efficient neural network acceleration with 3d memory. Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems, p.751-764.

Genc H, Kim S, Amid A, et al., 2021. Gemmini: Enabling systematic deep-learning architecture evaluation via full-stack integration. Proceedings of the 58th Annual Design Automation Conference (DAC).

Guo C, Xue F, Leng J, et al., 2024. Accelerating sparse dnns based on tiled gemm. *IEEE Transactions on Computers*, .

Hoefler T, Alistarh D, Ben-Nun T, et al., 2021. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1-124.

Inayat K, Muslim FB, Iqbal J, et al., 2023. Power-intent systolic array using modified parallel multiplier for machine learning acceleration. *Sensors*, 23(9):4297.

Inayat K, Muslim FB, Mahmood T, et al., 2024. Fpga-assisted design space exploration of parameterized ai accelerators: A quickloop approach. *Journal of Systems Architecture*, :103260.

Jouppi NP, Young C, Patil N, et al., 2017. In-datacenter performance analysis of a tensor processing unit. Proceedings of the 44th annual international symposium on computer architecture, p.1-12.

Kim B, Lee S, Trivedi AR, et al., 2020. Energy-efficient acceleration of deep neural networks on realtime-constrained embedded edge devices. *IEEE Access*, 8:216259-216270.

Lai C, Zhang W, 2024. gem5-nvdla: A simulation framework for compiling, scheduling and architecture evaluation on ai system-on-chips. *ACM Transactions on Design Automation of Electronic Systems*, .

Lee J, Kim C, Kang S, et al., 2018. An energy-efficient unified deep neural network accelerator with fully-variable weight precision for mobile deep learning applications. Hot chips: a symposium on high performance chips, hot chips: a symposium on high performance chips.

Li W, Liu T, Xiao Z, et al., 2023. Tcader: a tightly coupled accelerator design framework for heterogeneous system with hardware/software co-design. *Journal of Systems Architecture*, 136:102822.

Loh J, Dudchenko L, Viga J, et al., 2024. Towards hardware supported domain generalization in dnn-based edge computing devices for health monitoring. *IEEE Transactions on Biomedical Circuits and Systems*, .

Lym S, Erez M, 2020. Flexsa: Flexible systolic array architecture for efficient pruned dnn model training. *arXiv preprint arXiv:200413027*, .

Moghaddasi I, Nam BG, 2024. Enhancing computation-efficiency of deep neural network processing on edge devices through serial/parallel systolic computing. *Machine Learning and Knowledge Extraction*, 6(3):1484-1493.

Moons B, De Brabandere B, Van Gool L, et al., 2016. Energy-efficient convnets through approximate computing. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), p.1-8.

Moons B, Uytterhoeven R, Dehaene W, et al., 2017a. 14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi. 2017 IEEE International Solid-State Circuits Conference (ISSCC), p.246-247.

Moons B, Uytterhoeven R, Dehaene W, et al., 2017b. Dvafs: Trading computational accuracy for energy through dynamic-voltage-accuracy-frequency-scaling. Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, p.488-493.

Muslim FB, Inayat K, Khan S, 2024. Lpchisel: Automatic power intent generation for a chisel-based asic design. *Computers and Electrical Engineering*, 115:109115.

Park M, Hwang S, Cho H, 2024. Bird: Bi-directional input reuse dataflow for enhancing depthwise convolution performance on systolic arrays. *IEEE Transactions on Computers*, .

Qamar A, Muslim FB, Iqbal J, et al., 2017. Lp-hls: Automatic power-intent generation for high-level synthesis based hardware implementation flow. *Microprocessors and Microsystems*, 50:26-38.

Ryu S, Kim H, Yi W, et al., 2019. Bitblade: Area and energy-efficient precision-scalable neural network accelerator with bitwise summation. Proceedings of the 56th Annual Design Automation Conference 2019, p.1-6.

Seshadri K, Akin B, Laudon J, et al., 2022. An evaluation of edge tpu accelerators for convolutional neural networks. 2022 IEEE International Symposium on Workload Characterization (IISWC), p.79-91.

Song J, Cho Y, Park JS, et al., 2019. 7.1 an 11.5 tops/w 1024-mac butterfly structure dual-core sparsity-aware neural processing unit in 8nm flagship mobile soc. 2019 IEEE international solid-state circuits conference-(ISSCC), p.130-132.

Xu R, Ma S, Wang Y, et al., 2021a. Configurable multi-directional systolic array architecture for convolutional neural networks. *ACM Transactions on Architecture and Code Optimization (TACO)*, 18(4):1-24.

Xu R, Ma S, Wang Y, et al., 2021b. Heterogeneous systolic array architecture for compact cnns hardware accelerators. *IEEE Transactions on Parallel and Distributed Systems*, 33(11):2860-2871.

Xu R, Ma S, Guo Y, et al., 2023. A survey of design and optimization for systolic array-based dnn accelerators. *ACM Computing Surveys*, 56(1):1-37.

Yüzügüler AC, Sönmez C, Drumond M, et al., 2023. Scale-out systolic arrays. *ACM Transactions on Architecture and Code Optimization*, 20(2):1-25.