Energy Efficient Computing with Heterogeneous DNN Accelerators

MD Shazzad Hossain and Ioannis Savidis
Integrated Circuits and Electronics Laboratory, Drexel University, U.S.A.
Email: {msh89, is338}@drexel.edu

***Abstract*—The exploration of custom deep neural network (DNN) based accelerators for highly energy constrained edge devices with on-device intelligence is gaining traction in the research community. Despite the superior throughout and performance of custom accelerators as compared to CPUs or GPUs, the energy efficiency and versatility of state-of-the-art DNN accelerators is constrained due to the limited scope of monolithic architectures, where the entire accelerator executes only one model at any given time. In this paper, a multi-voltage domain heterogeneous DNN accelerator architecture is proposed that simultaneously executes multiple models with different power-performance operating points. The proposed architecture and circuits are evaluated with SPICE simulation in a 65 nm CMOS technology. The simulation results indicate that the proposed heterogeneous architecture improves the energy efficiency to 2.04 TOPS/W, while the conventional monolithic and single voltage domain architecture exhibits an energy efficiency of 0.0458 TOPS/W. In addition, the total power consumption of the accelerator SoC is reduced to 1.34 W as compared to the 3.72 W consumed by the baseline architecture when all multiply-and-accumulate (MACs) units operate at a voltage of 0.45 V.**

I. Introduction

On device artificial intelligence (AI) is a primary driving force for edge devices. Recent advances in deep neural network (DNN) models and DNN accelerators (customized hardware architectures optimized for DNN inference) have provided significant improvement in incorporating intelligence into ubiquitous edge devices, which are designed to meet stringent energy efficiency requirements [1]. The use of edge devices for applications including computer vision, augmented reality (AR), face recognition, image processing, and speech applications [1] require DNNs with variable specifications. The state-of-the-art DNN models customized for resource constrained edge devices are capable of inference with as little as 2-bit arithmetic [2], while training is demonstrated with as low as 4-bit arithmetic [3]. As the bit precision is reduced, the execution of both inference and training become more feasible on edge devices. However, the progress on the hardware level implementations of optimized DNNs has not sufficiently progressed as compared with the model and algorithmic breakthroughs made by the research community due to a) the lack of efficient circuits and architectures implementing the DNNs and b) the higher power consumption resulting from the large number of computations.

Regardless of the model, application, and hardware architecture, all DNN accelerators require a sufficiently large data set, where the data is primarily categorized into three types: input activation, output activation, and weights (or filters). The DNN accelerators are efficient in performing convolution operations on an array of processing elements (PEs), where each PE is composed of a multiply-and-accumulate (MAC) unit and local memories [4]. However, the DNN accelerators are composed of monolithic PEs (homogeneous accelerator architecture where all PEs are tasked with executing a single model at any given time), and all PEs are tied to a single power domain [4–6]. The implementation of a monolithic DNN architecture limits any improvement in the energy efficiency of the circuit as i) all PEs are allocated to a given model regardless of the actual hardware requirements of the executing model and ii) the same supply voltage is applied to all PEs regardless of the specific latency and throughput requirements of the executing model.

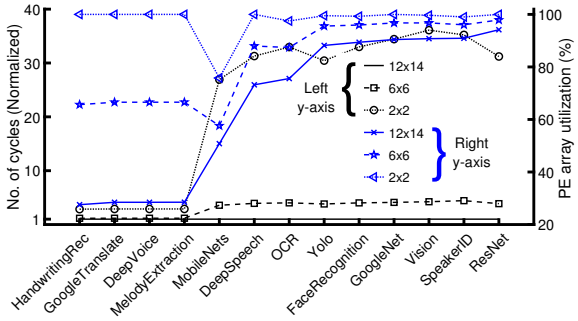


Fig. 1: Characterization of the PE utilization and the average number of cycles to execute a set of DNN models. Array sizes of 12×14, 6×6, and 2×2 are analyzed.

In this paper, a heterogeneous and multi-voltage domain DNN accelerator architecture is proposed, where multiple PEs are grouped to form N_{SA} number of sub-arrays that perform simultaneous execution of N_{SA} number of models. The primary contributions of this paper include

- a design space exploration that includes analysis of the computational demand and execution time requirements of the DNN models and any sub-layers needed by each model, and
- a novel multi-voltage domain heterogeneous DNN accelerator architecture, where each PE sub-array is connected to an independent voltage domain.

II. Heterogeneous DNN Accelerator with Multiple Voltage Domains

The development of an optimized DNN accelerator is an ongoing research effort [5, 6]. Several custom architectures were proposed in the past five years for DNN accelerators that provide improved performance, throughput, and energy efficiency as compared to CPU, GPU, and FPGA based implementations [5, 6]. However, the implementation of a DNN accelerator that offers both energy efficiency and performance across a diverse set of DNN models is a challenge as the architecture of DNN accelerators is often fixed and only optimized for a sub-set of DNN models [1]. The primary limitations of monolithic DNN accelerators are discussed in Sections II-A and II-B. The proposed multi-voltage domain DNN accelerator architecture is described in Section II-C.

A. Limited Benefit of Monolithic DNNs Accelerators

The computational requirements, on-chip memory size, and memory bandwidth of DNN accelerators vary by multiple

orders of magnitude when implementing different neural networks as well as across layers within a given neural network [6]. Maintaining a high energy efficiency when implementing a monolithic DNN accelerator is a challenge as a given dataflow does not map diverse DNN models and layers of a given model optimally to the available hardware resources. In this paper, an Eyeriss-like architecture is characterized using a cycle accurate neural processing unit (NPU) simulator [7]. The proposed architecture is analyzed across a set of neural network models and for three PE array sizes (12×14, 4×6, and 2×2). A weight stationary (WS) dataflow is applied for all array sizes and models. The average utilization of the PE array and the average number of cycles required to complete execution of a diverse set of neural network models used for applications that include vision, object detection, and speech recognition is characterized, with results as shown in Fig 1. The number of execution cycles is normalized to the number of cycles required by the 12×14 array. For the 12×14 PE array, the average number of cycles required to complete execution of HandwritingRec, GoogleTranslate, DeepVoice, MelodyExtraction, MobileNet, DeepSpeech, OCR, Yolo, FaceRecognition, GoogleNet, Vision, SpeakerID, and ResNet is, respectively, 0.245K, 2.84K, 2.48K, 3.627K, 206K, 1639K, 127K, 1735K, 423K, 174K, 1823K, 6007K, and 462K. The average PE utilization is less than 90% for most of the models when the PE array size is 12×14, while the PE utilization significantly increases for the smaller array sizes of 6×6 and 2×2. Underutilization of PE arrays across hundreds to thousands of cycles results in significant loss in energy due to leakage. An increase in the utilization of the PEs, therefore, results in a significant improvement in the total energy efficiency of the DNN accelerator. However, the reduction in the size of the array results in an increase in the average number of cycles needed to complete execution of the models by 1.16× to 4.43× and 2.86× to 36× for, respectively, arrays of size 6×6 and 2×2. The overall power-performance trade-off is, therefore, improved when the different models and layers are optimally mapped to a heterogeneous PE sub-array.

State-of-the-art edge devices execute multiple applications that concurrently run in the background. As an example, edge devices performing augmented reality require concurrent execution of object detection, speech recognition, pose estimation, and hand tracking [1]. In addition, due to the increasing complexity and the greater variety of DNN based workloads executed on edge devices, dynamic resource allocation is required [1]. Traditional DNN accelerators with monolithic architectures that are optimized to efficiently execute only a sub-set of models are, therefore, not well suited for current applications that require the execution of a diverse set of DNN models. Recent research proposed flexible and heterogeneous accelerators, where the heterogeneous DNN accelerators are best suited to improve the performance and energy efficiency of edge devices simultaneously running a diverse set of DNN models [1]. The heterogeneous DNN accelerators are composed of multiple sub-arrays of PEs each optimized for different layer shapes and operations [1]. Each sub-array of PEs is mapped to a dataflow that maximizes the resource utilization and improves the overall power-performance trade-off.

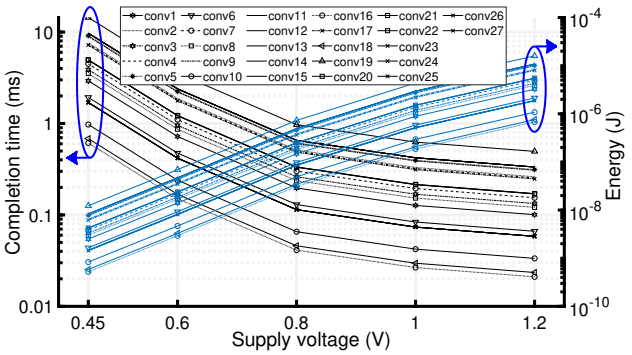


Fig. 2: A monolithic 12×14 PE array implementing the 27 convolutional layers of the MobileNets model when operating at each of five different supply voltages is characterized for energy consumption and completion time of each layer.

In addition, monolithic DNN accelerators, where all PEs share a common power domain, limit any improvement in energy efficiency provided by techniques such as fine-grained dynamic voltage scaling, power gating, adaptive voltage scaling, and the use of multiple voltage domains as all PEs are connected to a single supply voltage [1, 4, 7, 8]. For example, recently, an inference processor is implemented for improved energy efficiency, where all of the PEs are operated at a near-threshold voltage of 0.4 V for the entire execution of the DNN [5]. While power consumption is lower at 0.4 V, the operating frequency (60 MHz) is also significantly reduced [5], which limits the inference processor to the implementation of only a sub-set of DNN models. The highly constrained monolithic architecture is, therefore, not well suited for the execution of a diverse set of DNN model as the throughput, energy efficiency, and execution time of the accelerator are not dynamically adjustable.

The operating frequency of most of the state-of-the-art DNN accelerators is limited by the memory bandwidth despite the opportunity of running the computational units at much higher frequencies [4, 5]. For example, the Eyeriss accelerator implemented on a 65 nm CMOS process operates at a clock frequency of 200 MHz, where each PE consists of either a 16 bit MAC or two 8 bit MACs [4]. In addition, an inference processor operates at a 120 MHz frequency when set to a supply voltage of 0.7 V in a 65 nm CMOS technology [5]. Therefore, there are opportunities to improve the overall system-level energy efficiency by operating the computational units (MACs) at a lower supply voltage, while operating memory at a higher supply voltage.

B. Characterization of a Monolithic Accelerator Architecture Using the MobileNets Model

A monolithic accelerator architecture is characterized with the MobileNets model using a cycle accurate simulator of a neural processing unit [7, 9]. The number of MAC operations required to execute each of 27 convolutional layers of the MobileNets model is determined through simulation, which is then used to calculate the energy consumption of each layer of the neural network. The energy consumed per layer is characterized for five voltages ranging from 0.45 V to 1.2 V, with results as shown in Fig. 2 (right y-axis). The number

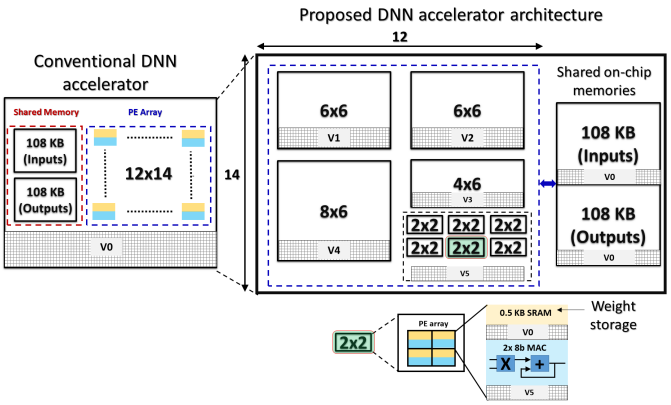


Fig. 3: The proposed multi-voltage domain heterogeneous DNN accelerator architecture implemented on Eyeriss based architecture.

of MAC operations is directly proportional to the completion time of each layer as shown in Fig. 2 (left y-axis), where the completion time at a given voltage is calculated based on the minimum cycle time of the two 8 bit MACs. For example, Conv27, the most computationally intensive layer of the MobileNets model, requires 3282 MAC operations when executed on the 12×14 PE array. The energy consumption (completion time) of Conv27 is 15.96 μJ (0.5 ms), 4.37 μJ (0.63 ms), 0.73 μJ (0.97 ms), 0.07 μJ (3.6 ms), and 0.01 μJ (14.4 ms) when operating at a supply voltage of, respectively, 1.2 V, 1 V, 0.8 V, 0.6 V, and 0.45 V. Conversely, Conv24 is the least computationally intensive layer requiring the execution of only 139 MAC operations. The resulting energy consumption and completion time of the Conv24 layer is respectively, 0.68 μJ and 0.021 ms when operating at a supply voltage of 1.2 V. Among the 27 convolutional layers, the standard deviation in the number of executed MAC operations is 845. Therefore, there is significant variation in the required computational resources and completion time across multiple layers of the neural network when characterizing for even a single model (MobileNets). The variation further increases with multiple models as discussed in Section II-A. Mapping the execution of the models to a heterogeneous PE array based on the required number of MAC operations and latency, therefore, provides benefit.

C. Proposed Heterogeneous DNN Accelerator Architecture

A multi-voltage domain heterogeneous DNN accelerator architecture is proposed to address the limitations of a monolithic DNN accelerator. The proposed architecture is shown in Fig. 3, where the accelerator is composed of multiple PE sub-arrays each operating in separate voltage domains, as opposed to established topologies that implement one large PE array with a single voltage domain as shown for the conventional DNN accelerator of Fig. 3. The inputs and outputs are stored in separate global on-chip memory, where the total size of each memory block is 108 KB. The weights required by each layer of the neural network are stored in on-chip memory of 0.5 KB blocks within each PE, where the total memory to store weights for 168 PEs is 84 KB.

The total number of PEs, number of MACs per PE, activation memory, and size of on-chip memory per PE are similar to that required by the Eyeriss architecture [4]. However, unlike

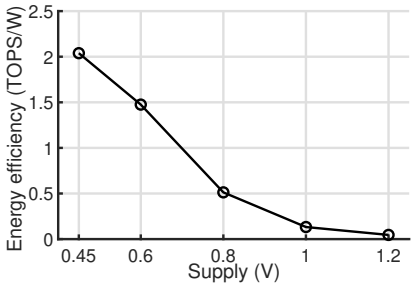


Fig. 4: Characterization of the energy efficiency across five voltage domains of an accelerator with a 12×14 PE array.

the Eyeriss architecture, the PEs of the proposed architecture are clustered into multiple sub-arrays each within independent voltage domains. Therefore, each PE sub-array executes a separate DNN model with an optimized performance-energy operating point.

III. Evaluation Framework

The proposed heterogeneous DNN accelerator, which consists of a total of 168 PEs and 216 KB of on-chip SRAM memory, is implemented and characterized through SPICE simulation in a 65 nm CMOS technology. Six voltage domains (V0 to V5) are implemented in Fig. 3 for the selection of the optimal power-performance point. Two configurations are considered to characterize the throughput and energy efficiency of the proposed accelerator. The baseline configuration includes a single voltage domain V0 set to 1.2 V that provides current to both the memory and MAC units. The proposed configuration includes two independent voltage domains, one for the on-chip memory and the other for the MAC units, which are set to, respectively, 1.2 V and 0.45 V. Therefore, the voltage of V0 is set to 1.2 V, while the voltage of V1, V2, V3, V4, and V5 are set to 0.45 V.

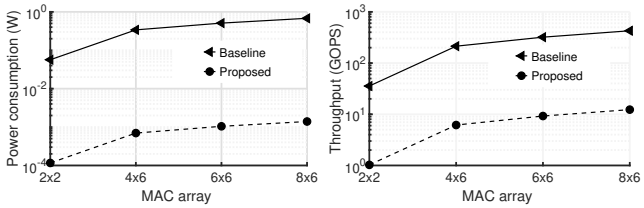
Each PE contains 1) two fixed-point 8-bit MACs with two-stage pipelines that produce two multiply-and-accumulate results per cycle and 2) 0.5 KB of on-chip SRAM. Each 8-bit MAC unit consists of a radix-4 booth multiplier and a carry look-ahead adder. The characterization of the proposed heterogeneous DNN accelerator is performed with 168 PEs that are clustered into ten sub-arrays: a) one 8×6 sub-array with a throughput of 3.67 giga operations per second (GOPS), b) two 6×6 sub-arrays each with a throughput of 2.75 GOPS, c) one 4×6 sub-array with a throughput of 1.84 GOPS, and d) six 2×2 sub-arrays each with a throughput of 0.306 GOPS. The number of sub-arrays and the size of each sub-array are chosen such that the different throughput requirements of the DNN models and corresponding layers are met.

IV. Characterization of Energy Efficiency and Throughput of the Proposed Architecture

The energy efficiency of an accelerator consisting of 168 PEs and 336 MAC units that are evenly distributed, is characterized in tera operations per second/per watt (TOPS/W), for the five voltages of 1.2 V, 1 V, 0.8 V, 0.6 V, and 0.45 V. The results from the analysis of the energy efficiency are shown in Fig. 4. The energy efficiency increases as the supply voltage is reduced, where the energy efficiency of the entire accelerator is 44.5× greater at 0.45 V (2.04 TOPS/W) than at 1.2 V (0.0458 TOPS/W). Note that the total power consumption and delay

of the MAC arrays is considered when calculating the energy efficiency for each voltage.

The total power consumption and the throughput of the MAC arrays are characterized using the baseline and the proposed architecture with results shown in Fig. 5. The power consumption of the baseline is 487.2× that of the proposed technique as shown in Fig. 5(a). The throughput of the baseline is 35× that of the proposed technique as shown in Fig. 5(b). The relative difference in both the total power consumption and energy efficiency scales with the size of the sub-array when comparing the baseline and the proposed configurations.



(a) (b)

Fig. 5: Characterization of the baseline and proposed technique for a) total power consumption and b) throughput.

TABLE I: Characterization of the total power consumption of 2×2 PE sub-arrays, where a single PE is composed of a 0.5 KB memory and two 8-bit MACs.

	Total power consumption (W)	
	One 2×2 sub-array	Six 2×2 sub-array
Baseline	0.066	0.393
Propose	0.009	0.054

V. Characterization of the Total Power Consumption

The total power consumption of a 2×2 sub-array and a group of six 2×2 sub-arrays (as shown in Fig. 3) are listed in Table I. Each 2×2 sub-array of PEs consists of 8 MAC units and a total of 2 KB SRAM memory. The total power consumption of one 2×2 sub-array is 66 mW and 9 mW for, respectively, the baseline and the proposed architecture. The power consumption increases proportionally to the array size as the characterization of the topology with six 2×2 sub-arrays resulted in a total power consumption of 393 mW and 54 mW for the baseline and the proposed architecture, respectively. Operating the MAC units of one 2×2 (six 2×2) sub-array(s) at 0.45 V, therefore, results in a 57 mW (339 mW) reduction in the total power consumption. The six 2×2 sub-arrays of PEs constitute 14.3% of all PEs (24/168), where all MACs within the 2×2 sub-array operate at 0.45 V and the on-chip memories operate at 1.2 V. The total power consumption is characterized for four different percentages of the MAC units operating at a supply voltage of 0.45 V (14%, 25%, 50%, and 100%), while the remaining units are set to 1.2 V for each case, with results as shown in Fig. 6. The total power consumption includes the on-chip weight memory (84 KB), MAC units within each of the 168 PE, and the 216 KB of on-chip global activation memory. The power consumption is significantly reduced as more MACs operate at a supply voltage of 0.45 V, where the total power consumption of the accelerator SoC is reduced to 0.68× (2.53 W) and 0.36× (1.34 W) that of the baseline (3.72 W) for, respectively, 50% and 100% MAC operation at 0.45 V.

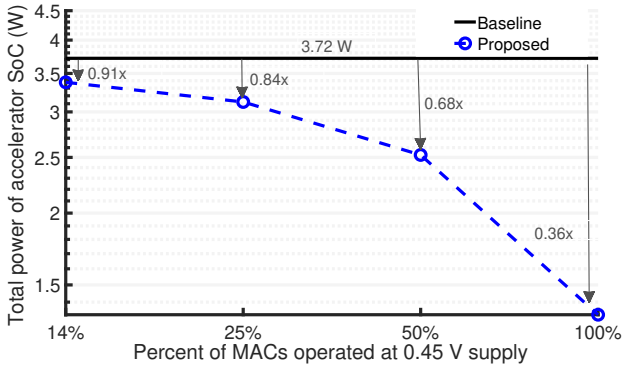


Fig. 6: Analysis of the total power consumption of the accelerator for four different percentages of the MAC units operating at a 0.45 V, while the remaining units operate at 1.2 V.

VI. Conclusions

A design space exploration is performed that indicates that a monolithic accelerator is not optimized for use with multiple DNN models. A heterogeneous multi-voltage domain DNN accelerator is proposed that implements multiple PE sub-arrays for the concurrent execution of different neural network models with varying throughput requirements. The proposed architecture includes MAC units operating at 0.45 V and on-chip memory operating at 1.2 V. For workloads that do not require high throughput, the energy efficiency is improved by 44.5× (2.04 TOPS/W) when comparing the proposed architecture to a monolithic DNN topology (baseline) that includes memory and MAC units both set to a single voltage of 1.2 V. Operating the MAC units of one 2×2 (six 2×2) sub-array(s) at 0.45 V reduces the total power consumption by 57 mW (339 mW). In addition, operating the memory and MAC units at separate voltages results in a reduction of the overall power consumption of the SoC to 0.36× that of the baseline.

References

[1] H. Kwon, L. Lai, T. Krishna, and V. Chandra, "HERALD: Optimizing heterogeneous DNN accelerators for edge devices," *arXiv preprint arXiv:1909.07437*, pp. 1–13, December 2020.

[2] J. Choi, S. Venkataramani, V. Srinivasan, K. Gopalakrishnan, Z. Wang, and P. Chuang, "Accurate and efficient 2-bit quantized neural networks," *Proceedings of the Systems and Machine Learning (SysML) Conference*, pp. 1–12, April 2019.

[3] X. Sun, N. Wang, C. Chen, J. Ni, A. Agrawal, X. Cui, S. Venkataramani, K. El Maghraoui, V. V. Srinivasan, and K. Gopalakrishnan, "Ultra-low precision 4-bit training of deep neural networks," *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1796–1807, December 2020.

[4] Y. Chen, T. Yang, J. Emer, and V. Sze, "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, Vol. 9, No. 2, pp. 292–308, April 2019.

[5] J. Sim, S. Lee, and L. Kim, "An energy-efficient deep convolutional neural network inference processor with enhanced output stationary dataflow in 65-nm CMOS," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 28, No. 1, pp. 87–100, September 2019.

[6] P. Jokic, S. Emery, and L. Benini, "Improving memory utilization in convolutional neural network accelerators," *IEEE Embedded Systems Letters*, pp. 1–4, July 2020.

[7] A. Samajdar, Y. Zhu, P. Whatmough, M. Mattina, and T. Krishna, "Scale-sim: Systolic CNN accelerator," *arXiv preprint arXiv:1811.02883*, pp. 1–11, February 2019.

[8] M. S. Hossain and I. Savidis, "Dynamic idle core management and leakage current reuse in MPSoC platforms," *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 49–54, August 2020.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, pp. 1–9, April 2017.