VWA: Hardware Efficient Vectorwise Accelerator for Convolutional Neural Network

Kuo-Wei Chang, Tian-Sheuan Chang, Senior Member, IEEE

Abstract—Hardware accelerators for convolution neural networks (CNNs) enable real-time applications of artificial intelligence technology. However, most of the existing designs suffer from low hardware utilization or high area cost due to complex dataflow. This paper proposes a hardware efficient vectorwise CNN accelerator that adopts a $3 \times 3$ filter optimized systolic array using 1-D broadcast dataflow to generate partial sum. This enables easy reconfiguration for different kinds of kernels with interleaved input or elementwise input dataflow. This simple and regular data flow results in low area cost while attains high hardware utilization. The presented design achieves 99%, 97%, 93.7%, 94% hardware utilization for VGG-16, ResNet-34, GoogLeNet, and Mobilenet, respectively. Hardware implementation with TSMC 40nm technology takes 266.9K NAND gate count and 191KB SRAM to support 168GOPS throughput and consumes only 154.98mW when running at 500MHz operating frequency, which has superior area and power efficiency than other designs.

Index Terms—Convolution neural networks (CNNs), hardware design, accelerators

I. INTRODUCTION

Deep convolution neural networks (DCNNs) have been widely used in computer vision tasks, such as recognition [1-5], detection [6-10], and autonomous vehicles during recent years for its significant improvement over traditional approaches. However, the computation of CNNs demands a lot of multiplications and accumulations (MACs) and millions of data amount per layer that prohibits its wide usage in real-time applications. Thus, hardware acceleration for DCNNs is demanded to satisfy the computation and bandwidth requirements under the real time constraint.

Various hardware accelerators have been proposed recently [11-23, 33-38]. [11]-[15] minimize the communication traffic by a tiling strategy and internal buffer. However, the modern DCNNs are too large to store all weights into the on-chip memory. [16] adopts a spatial array architecture and row stationary data flow that helps maximize data reuse but leads to high processing element (PE) cost due to large local storage. In addition, its fixed PE configuration results in low hardware utilization. Its improved version [17] has higher hardware utilization and supports sparse CNNs by clusters to increase throughput, but needs large area overhead. [18] proposes a systolic array architecture with full reconfigurations for different convolutional kernels. It achieves high hardware utilization but needs significant PE area cost due to complex control overhead. The design in [19] is precision scalable to achieve low power consumption under different application scenarios. [20] uses a filter-type structure that can be reconfigurable to fit different needs but this filter-like data flow limits the flexibility of reconfigurations. The streaming architecture in [21] blackeuces data movement through filter reuse and input reuse to optimize energy efficiency, but it has low hardware utilization beyond optimized $3 \times 3$ architectures.

The systolic array architecture in [22] is also a reconfigurable design to fit different convolution kernels but with a propagated input data flow that results in long latency and low hardware utilization. [23] proposes an accelerator for image segmentation which supports dilated and transposed convolution. In addition, it cuts down the blackandant zero computations for higher throughput. [33] provides near non-blocking communication via reconfigurable links with high bandwidth to support efficient irregular dataflow mapping and provide high utilization of computing units. [34] proposes a low precision design that uses heterogeneous representation and significant bits securing encoding which can exploit the advantage of both efficiency of fixed-point and flexibility of floating-point to optimize their PE architecture and consume less power and area. [35] adopts different combinations of input, weight and output data flow to optimize their design. The data flow in [36] inspiblack from the fully-connected layer can accelerate convolutional processes and parametrize design from low memory accesses to low latency systems. [37] is a computing in memory design that adopts SRAM array for vector matrix multiplication for low power deep learning applications. [38] implements convolutions of different kernel sizes with multiple parallel fast FIR based convolution units. Although existing designs can support most of kernel sizes and stride lengths, they suffer from low hardware utilization or high hardware cost due to constrained data flow or complex PE structures for such reconfiguration requirements.

To achieve high hardware utilization with low area overhead, this paper proposes a hardware efficient vectorwise accelerator based on a PE array for CNN. This design adopts broadcasted vectorwise input and weight for the PE array to avoid local weight storage. The array is partitioned into blocks and optimized for the widely used $3 \times 3$ convolutions but also easily reconfigublack for other kernel size due to the broadcasted data flow. The PE array has high hardware utilization even for the non-unit stride case by the interleaved input.
This design can also support $1 \times 1$ convolution commonly used in the low complexity network with the elementwise input dataflow. Even with above versatile reconfigurations, the proposed design still has high hardware efficiency comparable to other designs.

The rest of the paper is organized as following. Section II gives a short overview of the CNN algorithm. Section III presents the proposed architecture. Section IV shows the details of the data flow for all kinds of convolutional filters supported in our architecture. The implementation results and comparisons are shown in Section V. Finally, we conclude this paper in Section VI.

II. OVERVIEW OF CNN

Fig. 1 shows a typical DCNN model, called AlexNet [26], that classifies a 224×224 RGB image into one of 1000 categories. A typical CNN structure consists of the convolutional layers, pooling layers, fully-connected layers, and the activation function. The convolutional layers transform images into highly abstract representations called feature maps. The pooling layers execute down-sampling operations to decrease feature map size and improve the translational invariance of features. In the state-of-art networks, the pooling layers could be replaced by layers of $3 \times 3$ convolution with stride two. The commonly used activation function like rectified linear unit (ReLU) adds non-linearity after the convolutional layers or the fully-connected layers to improve feature representations. The fully-connected layers classify the image based on those extracted features, which is a pure neural network with a lot of weight numbers.

The convolutional layer applies a filter kernel to the whole image, which can be defined as

$$O(b, c, r, w) = Act\left(\sum_{u=0}^{U-1} \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} F(b, u, Sr + i, Sw + j) \times W(c, u, i, j) + B(u)\right)$$

where $O$, $F$, $W$, and $B$ are the matrices of output feature maps, input feature maps, weight, and bias, respectively. In which, $Act$ is the activation function, such as ReLU, $b$ is the image batch size, $c$ is the number of filters, $r$ is the output row size, $w$ is the output column size, $u$ is the input channel, $i$ is the weight row size, $j$ is the weight column size, and $S$ is the stride length. For this whole image operation, the computation needs $c \times u \times i \times j \times r \times w \times b$ MACs for one layer. The computation cost of the whole model is increased significantly with modern deeper and wider models, which will be over tens or hundreds of giga or tera operations per second for models like ResNet and GoogLeNet [39].

III. SYSTEM ARCHITECTURE

A. Overview

Fig. 2 shows the proposed system architecture, which consists of eight PE blocks, three-stage accumulators and a boundary data SRAM buffer for convolution computation, a post-processing module for ReLU, batch-normalization and pooling, and multi-bank input and weight SRAM buffers for continuous data access from external memory. The weight data buffer is a ping-pong buffer, which is divided into eight banks for eight PEs, respectively. The input buffer is divided into 24 input data banks.
banks for eight PEs. Each PE block will connect to three input SRAM banks to support different filter sizes.

The proposed design will first load a tile of input map and weight data to the local SRAM buffers for data reuse. Each PE block will use one channel of input map and one weight filter to compute partial convolution results. With eight PEs, we will be able to process eight input channels and eight filters at the same time. The partial convolution results from all PE blocks will be stroblocak and accumulated locally through the three stage accumulators to avoid external memory access. During the above accumulation, partial results at the top and bottom boundary of tiles will be also stroblocak in the boundary buffer to avoid repeated computation at the tile boundary. The final convolutional result will be processed by the post-processing module for ReLU, batch-normalization, and pooling operations, and then stroblocak to external memory.

B. PE Blocks

Fig. 3 illustrates the 7×3 MAC array design of one PE block. The PE block has 3×7 inputs broadcasted horizontally and three weights broadcasted vertically that are multiplied together and then summed along horizontal direction for an 1×1 filter or diagonal direction for other filter sizes. To support different convolution filters, each input row has three inputs so that each MAC can select the desired input via a 3-to-1 multiplexer. This design chooses a three-column width for weights to optimize computations of 3×3 convolutions and avoid long critical path. With this proposed data flow, we can maximize hardware utilization while keep data flow regular for lower area cost.

Beyond the above block level reconfiguration, the proposed eight PE blocks can also be reconfigublack as two configurations, (8, 7, 3): 8 blocks, and 7 rows by 3 columns per block, or (4, 14, 3): 4 blocks, and 14 rows by 3 columns per block that combines two PE blocks into one as in Fig. 2. This array level reconfiguration can support different input map sizes while maximize hardware utilization. For example, the PE array is reconfigublack as (4, 14, 3) in the first layer of the models for image recognition, such as AlexNet, VGGNet, and ResNet because the input map only consists of three channels with R, G, and B. But the PE array can be reconfigublack as (8, 7, 3) in other layers due to wider channel numbers.

IV. DATA FLOW OF DIFFERENT KERNELS

The proposed accelerator is designed to optimize computations of 3×3 filters stride one with a vectorwise scheduling due to its wide usage. For larger filter size, we adopt kernel decomposition method [20] so that this design only needs extra support of 4×4, and 5×5 filters. In addition, with the popularity of low complexity models, this design also supports 3×3 with stride two by interleaved input, depthwise convolution by reconfigurable accumulators and 1×1 filter by elementwise input dataflow for higher hardware utilization.

A. Basic 3×3 Convolution with Unit Stride

Fig. 4 shows the data flow of a PE block to compute an example with 5×6 input tile and a 3×3 filter kernel as in Fig. 3. This example case assumes a 5×3 MAC array for clarity. In which, the blocks with the same color are the partial results for the same convolution output. Above computation also includes tile boundary cases on top (e.g. \( a_0, b_0, c_0, d_0, e_0, f_0 \)) and bottom (e.g. \( a_5, b_5, c_5, d_5, e_5, f_5 \)) which will be stroblocak in the global boundary buffer. For a 3×3 convolution, one column vector of input is broadcasted horizontally along the MAC.
Fig. 6. Dataflow chart for the example in Fig. 5 with 5x3 MACs. In which, the same color elements belong to the same filter output, which will be summed together in a PE at the same cycle and then accumulated later in the accumulator at different cycles, e.g. black elements at t = 2, 5, and 9 accumulated for $ob_2$.

Fig. 7. The computational order for $3 \times 3$ convolutions with unit stride, where $oa$ to $od$ are the column vector of the output. $a$ to $f$ are the column vector of the input. $wa$ to $wc$ are the column vector of the weight. Numbers 1-12 represent cycle numbers in Fig. 6.

Fig. 8. The computational order for $4 \times 4$ convolutions with unit stride, where $wa_0-2$ is a vector with $wa_0, wa_1, wa_2, wb_0-2$ is a vector with $wb_0, wb_1, wb_2$, and so on. $oa = oa_0 + oa_1$, $ob = ob_0 + ob_1$, $oc = oc_0 + oc_1$.

Fig. 9. The computational order for $5 \times 5$ convolutions with unit stride, where $wa_3-4$ is a vector with $wa_3, wa_4$, $wb_3-4$ is a vector with $wb_3, wb_4$, and so on. $oa = oa_0 + oa_1$, $ob = ob_0 + ob_1$.

Fig. 10. An example of $3 \times 3$ convolutions/stride 2 with $5 \times 5$ input to generate $2 \times 2$ output.

This vectorwise scheduling enables simple dataflow for high hardware utilization and low area cost. For $3 \times 3$ depthwise convolution, the output of the first stage of the accumulators will be the desiblack result.

The overall vectorwise scheduling is shown in Fig. 6 which takes 12 cycles to complete the example case. The concept of this scheduling is to decompose a convolution into sums of...
MAC results by an input column and a weight column as shown in Fig. [7]. Each MAC result of an input column and a weight column will take one cycle to complete. Thus, a 3×3 filter computation will need three cycles. However, these three cycles are not continuous. We rearrange the computation order as shown in Fig. [7] so that the same input can be reused successively for the computation to save the power of input buffer SRAM. Besides, this reorder enables a simple continuous addressing of the input SRAM buffer to blackuce control overhead.

With the above computation, the speedup for the above example is \( \frac{\text{number of MACs}}{\text{cycles}} = \frac{180}{12} = 15 \), which is equal to the number of PEs \((3 \times 5 = 15)\). Thus, the hardware is fully utilized.

**B. 4×4 and 5×5 Convolution with Unit Stride**

Beyond 3×3 convolution with unit stride, our proposed architecture also supports 4×4 and 5×5 convolutions to implement different filter sizes in current popular models. The 4×4 and 5×5 convolutions will take two PE blocks. As shown in Fig. [8] and Fig. [9] the first three elements (e.g. \( w_{a0} - w_{a2} \), \( w_{d0} - w_{d2} \) in 4×4, or \( w_{a0} - w_{a2} \), \( w_{e0} - w_{e2} \) in 5×5) of the filter columns will be at the first block and the remaining elements (e.g. \( w_{a3} - w_{d3} \) in 4×4, or \( w_{a3} - w_{a4} \), \( w_{e3} - w_{e3} \) in 5×5) will be at the second block with two or one column of unused MACs. The overall data flow is the same as the one in the 3×3 convolution but with properly weight assignments for 4×4 and 5×5. With this approach, the hardware utilization will be lower since part of the PE block will not be used. This design tradeoff can make hardware regular since these kernel
TABLE I: Per layer DRAM access amount for VGG-16.

| Layers | Input Access (MB) | Weight Access (MB) | Output Access (MB) | Total Access (MB) | Total Energy (mJ) | Tile Size (7x7n) |
|--------|------------------|--------------------|-------------------|------------------|-----------------|-----------------|
| 1      | 0.287            | 0.002              | 6.125             | 6.432            | 0.45            | 224             |
| 2      | 6.125            | 0.035              | 6.125             | 12.285           | 0.86            | 112             |
| 3      | 1.531            | 2.25               | 3.063             | 6.844            | 0.479           | 112             |
| 4      | 3.063            | 9                  | 3.063             | 15.125           | 1.059           | 56              |
| 5      | 0.766            | 4.5                | 1.531             | 6.797            | 0.476           | 56              |
| 6      | 1.531            | 18                 | 1.531             | 21.063           | 1.474           | 56              |
| 7      | 1.531            | 18                 | 1.531             | 21.063           | 1.474           | 56              |
| 8      | 0.383            | 9                  | 0.766             | 10.148           | 3.02            | 28              |
| 9      | 0.766            | 36                 | 0.766             | 37.531           | 2.627           | 14              |
| 10     | 0.766            | 36                 | 0.766             | 37.531           | 2.627           | 14              |
| 11     | 0.191            | 9                  | 0.191             | 9.383            | 0.65            | 14              |
| 12     | 0.191            | 9                  | 0.191             | 9.383            | 0.65            | 14              |
| 13     | 0.191            | 9                  | 0.191             | 9.383            | 0.65            | 14              |
| Total  | 17.322           | 159.805            | 25.84             | 202.967          | 14.208          |                 |

sizes are only used in the first input layer in most of the current network models.

C. 3×3 Convolution with Stride 2

For a 3×3 convolution with stride 2, a naive dataflow implies a 25% hardware utilization since only 25% of PE result is useful for the final result. To increase hardware utilization, we propose interleaved input as shown in Fig. 11. Thus, we add a multiplexer at the input data of multiplier in each PE as shown in Fig. 3 to support this. Fig. 11 shows the data flow of a PE block for an example with a 5×5 input, and 3×3 filer kernel with stride 2. Input data from two different input columns (e.g. \( a_0 - a_4, c_0 - c_4 \)) will be interleaved as the input of successive PE rows to increase hardware utilization. An example as shown in Fig. 10 with stride 2 only need 3 cycles to generate \( a_0c_0, a_0c_2, a_2c_0, \) and \( a_2c_2 \). The result will be accumulated at the accumulator as 3×3 convolution with unit stride but with interleaved results. Thus, the speedup is \( \frac{\text{number of MACs}}{\text{cycles}} = \frac{36}{3} = 12 \). However, the PE block will have some MACs unused due to interleaved input. So the utilization is \( \frac{\text{speedup}}{\text{number of PEs}} = \frac{12}{18} = 67\% \) in this example, which is lower than that in 3×3 convolutions with unit stride but higher than that with a naive mapping.

D. 1×1 Convolution

1×1 convolutions are getting popular in modern deep learning models. However, unlike other convolutions, the 1×1 convolution only has size one in the spatial dimension, which will result in much more output than 3×3 convolutions if we implement it naively on the current 3×3 convolution based PE blocks. This implies more parallel hardware needed for the following accumulators to keep PE array full utilized or the PE array stalled during the accumulation.

To solve above problems, we propose the elementwise input scheduling for 1×1 convolutions as shown in Fig. 12 that will have distinctive input elements in each multiplier. In this scheduling, each PE block will select their specific input feature map via a 3-to-1 multiplexer. With such input, the summation is now along the horizontal direction. Thus, we will have output number equal to block number = 8 for eight PE blocks. Each output from a PE block will have sum of three channel multiplications. In this flow, all PE outputs are just different parts of a channel output. Thus, we will accumulate them with the first two stages of the accumulator as shown in Fig. 13. The stage three is skipped in this case. With above flow, we can achieve high hardware utilization.
V. EXPERIMENTAL RESULT

A. Experiment Setup

This design has been implemented with the TSMC 40nm CMOS technology. The hardware power consumption is estimated by Synopsys PrimeTime PX. We also build a cycle based tool by Pytorch to properly evaluate hardware utilization, speedup and DRAM access under different models and PE array configurations. In addition, this tool also helps generate configuration information to configure the hardware controller. For the following analysis, we will use four famous models, VGG-16 [27], MobileNet [29], ResNet-34 [28], and GoogLeNet [30], trained on the ImageNet dataset as our test cases.

B. Implementation Result

Fig. 14 shows the chip layout and its implementation result. The core area is 1.25mm × 1.25mm with 191KB SRAM including 99KB input SRAM bank, 36KB weight SRAM bank, and 56KB global boundary data SRAM. The peak performance is 168 GOPS at 500MHz based on the technology library from the worst PVT corner, when the PE utilization is 100%. The power consumption is 154.98mW, measured by running VGG-16 convolutions based on the technology library from the best PVT corner at 0.99 V power supply.

C. Memory Access and Energy Consumption Analysis

The external memory access of this design is blacked by data reuse. The input is divided into configurable tile size with 7 × n size for the ImageNet dataset, where n is denoted as in Table I. The input SRAM buffer has size enough to store 7 × 6 tiles with 1024 input channels for commonly used models. The proposed design will first compute all channels of an input tile with all weight data to get the final output tile. During this computation, all input data, partial sum, and output will access local buffer instead of external DRAM. Only the weight data will be accessed from DRAM repeatedly when the input tiles are changed. So, external DRAM access can be formulated as follows:

$$\text{Input Access} = \text{input channel} \times \text{image size} \quad (2)$$

$$\text{Weight Access} = \text{number of input tiles} \times \text{output channel} \times \text{input channel} \times \text{weight size} \quad (3)$$

$$\text{Output Access} = \text{output channel} \times \text{output size} \quad (4)$$

The total energy is 70pJ/bit × total access for DDR3 DRAM [32]. As shown in Table II, there are no blacked accesses in input and output. The weight accesses are correlated with the number of tiles. With the configurable tile size, we can tradeoff between tile size and input channel to decrease the number of tiles and weight accesses.
Fig. [15]a) shows power breakdown of the logic circuit in the core. SRAM accounts for 62% of the total power. The combinational logic consumes 13% for MAC and accumulator computations. The sequential logic consumes only 20%, because of the lower register numbers in this data broadcasted design. This also results in lower power consumption in the clock network. Fig. [15]b) shows power breakdown of architecture elements in the core. The PE array accounts for 17% due to 168 MACs. The stage one of the accumulator occupies 14% because it keeps accumulation in most of the cycles and consists of eight parallel accumulators. The other stages of the accumulator occupies only 1% because they only operate after a valid stage one output.

D. Area Analysis

Fig. [16]a) shows area breakdown of the core. Without SRAM, PE array occupies 70% of total area due to 168 MACs while accumulator occupies 29%. Area of the controller is the lowest due to the simple and regular data flow. For the whole core area in Fig. [16]b), SRAM buffers occupies 78.6% due to large SRAM buffer for data reuse.

E. Analysis of Hardware Utilization and Performance

Fig. [17] to 20 show hardware utilization of each layer in VGG-16, Resnet-34, MobileNet, and GoogLeNet with two PE configurations: (8, 7, 3) or (4, 14, 3). In which, (3×3/s1 3|64) as shown in Fig. [17] to 20 represents 3×3 convolutions with unit stride, 3 input channels, and 64 output channels.

For VGG-16, it consists of all 3×3 convolutions and thus has 100% utilization for all layers except the first layer. The first layer only has 75% utilization for (4, 14, 3) since only 3 instead of 8 blocks are used to compute 3 channel input.

For ResNet-34, the first layer is 7×7 convolutions with stride 4, which can be decomposed into 4×4 convolutions with unit stride. However, 4×4 convolutions do not achieve 100% of hardware utilization. Thus, the first layer will have only 77% of utilization for the (4, 14, 3) case or 38% of utilization for the (8, 7, 3) case. Most of the middle layers achieve 100% of utilization for its 3×3 convolutions except for the layers of the 3×3 convolution with stride 2 that have only 80% of utilization. The utilization of the last layer of the 3×3 convolutions with stride 2 is dropped to 58% due to small 7×7 feature maps. The final few layers have 49% of utilization for the (4, 14, 3) case due to small feature map size, 7, in large PE rows, 14, of a block. To improve hardware utilization, we reconfigure the PE array to be (4, 14, 3) for the first layer, and (8, 7, 3) for the last few layers with the layer adaptive PE configuration, and thus have the best one from (8, 7, 3) and (4, 14, 3) cases.

For MobileNet, the layers with 3×3 stride 1 have 100% of utilization, 1×1 layers have 89% of utilization since we only process 8×3 = 24 or 4×3 = 12 channels at a time and cannot have full hardware utilization for the certain number of channels. The layer with 3×3 stride 2 has lower hardware utilization with the same reason as above. In this case, we can also use layer adaptive PE reconfiguration as above for higher utilization. In summary, with the proposed data flow, we can attain high utilization while keep structure simple.

For GoogLeNet, utilization of the most layers with 3×3 and 1×1 achieves near 100% of utilization due to our reconfigurable PE array for different convolution kernels. Only the first one and last few layers have lower utilization due to the same reasons as in MobileNet. However, with our configurable PE blocks, we can attain higher utilization for these layers.

Fig. [21](a) shows the overall hardware utilization in different models and PE configurations. The utilization of 4 blocks is always lower than 8 blocks because hardware utilization is only half when the size of the feature map is down-sampled to 7 (half of 14 rows) for ResNet-34, GoogLeNet, and MobileNet. VGG-16 is the only exception since its minimum feature map size is 14. The layer adaptive PE configuration has the best utilization. With the proposed data flow, the utilization can achieve 99%, 97%, 93.7%, and 94% for VGG-16, ResNet-34, MobileNet and GoogLeNet, respectively.

Fig. [21](b) shows the overall throughput in different models and PE configurations. These throughput numbers are proportional to the hardware utilization since our models have optimized for different convolutional types in these models.

F. Design Comparison

Table. II shows the implementation result and comparison with other designs. The proposed design has much lower area cost than other designs due to the simpler PE structure and regular local summation. The peak throughput can reach up to 168 TOPS. The area efficiency is 0.629 GOPS/KGE and 107.52 GOPS/mm², which is much higher than other designs. The power consumption is 154.98mW and power efficiency can reach up to 1.084 TOPS/W because of the simple and regular data flow. The power efficiency is better than most of other designs except for [19] and [34]. [19] uses lower bitwidth hardware to compute 30% - 60% sparse network at lower supply voltage to attain lower area cost or power consumption. [34] uses an optimized numerical representation for lower bitwidth hardware to achieve lower power consumption.

Table III shows a layer-by-layer performance comparison for AlexNet. For a fair comparison, we first list speedup ratio of the peak throughput between different designs and the proposed work. This can normalize the effects of different clock rate and PE numbers in each design, which implies a full hardware utilization case. Then, for each layer, if the compablack design has lower hardware utilization than the proposed one, the speedup ratio will become larger. Otherwise, the speedup ratio will be lower. By dividing speedup of the layers with speedup of the peak throughput, we will know the difference in hardware utilization between designs. Thus, compablack to [16], the hardware utilization of our design is 20.4% (3.139 - 2.5/3.139) higher. When compablack to [19], the hardware utilization of our design 1.95% (0.804 - 0.82/0.82) lower since our design has lower hardware utilization for CNN with low number of channels and non-3×3 convolutions at the first and second layers of AlexNet, respectively. However, this difference is quite small. With our 3×3 optimized based structure, we can have the simpler structure and lower cost
TABLE II: Implementation result and comparisons with other designs.

| Technology | Measurements     | Our work | 16  | 17  | 18  | 19  | 21  | 33  | 34  |
|------------|------------------|----------|-----|-----|-----|-----|-----|-----|-----|
|            |                  | 40nm     | 65nm| 65nm| 65nm| 28nm| 65nm| 65nm| 65nm|
| Supply Voltage (V) |            | 0.99     | 1.0 | 1.0 | 1.2 | 0.65 - 1.1 | 1.0 | -   | 1.2 |
| Precision (bits)   |            | 16 fixed | 16 fixed | 8-20 fixed | 16 fixed | 4, 8, 16 Dyna. | 16 fixed | 16 fixed | 10 | 16 fixed |
| Power (mW)        |            | 168      | 168 | 192 | 312 | -   | 168 | 168 | 734 | 128 | 64 |
| Area (mm²)        |            | -        | 67.2 | 153.6 | 204.8 | 102 - 408 | 152 | 67.2 / 149.6 | 25.6 | 23.4 |
| SRAM(KB)          |            | -        | 1.24 | - | - | - | - | - | - | - |
| Area eff. (GOPS/KGE) |         | 0.56 - 0.27 | 0.18 | - | - | - | - | - | - | - |
| Power eff. (TOPS/W) |        | 1.084 | 0.399 | 0.552 | 0.969 | 3.017 | 0.711 | - | 2.96 | 0.596 |

Notes:
- 1GMACS= 2 GOPS
- 4Technology scaling (Process 40nm)
- 5Core only size.
- 6Core only power.
- 730% - 60% sparsity, 3 - 4bits, and 76 GOPS at 0.65V.
- 8The area is shown in terms of the size of kido NAND2 gates (KGE).
- 9We take the theoretical performance to evaluate area efficiency fairly here.
- 10Chip size.
- 11Normalized power efficiency = power efficiency × (Process 40nm) × (Voltage, 0.99V)².
- 1216 bits precision and 76 GOPS at 1.1V.

Beyond the first two layers, the utilization of our design is 12% higher at other layers, which proves the efficiency of the proposed design.

Table IV shows another comparison for VGG-16. The utilization of optimized PE configurations is 470% higher than that in [16]. The reason is that these layers are our optimized 3×3 layers and have 100% of hardware utilization. The only exception is the first layer due to only three input channels. The large hardware utilization degradation at CONV1 to CONV7 in [16] is due to bad handling of the large feature map in their PE structure. On the contrary, our optimized 3×3, 4×4, 5×5, 3×3 stride 2, depthwise, and 1×1 convolutions. This flexible scheduling is enabled by layer adaptive PE configurations, interleaved input selection scheme, elementwise input, and reconfigurable accumulators. To keep PE hardware utilization high.

VI. CONCLUSION

This work was supported by Ministry of Science and Technology, Taiwan, under Grant 108-2634-F-009-005, 107-
TABLE III: Layer by layer comparison for AlexNet.

| PE number | Clock rate | Peak throughput (clock rate * PE) | Real throughput (GOPS) | CONV1 (ms) | CONV2 (ms) | CONV3 (ms) | CONV4 (ms) | CONV5 (ms) | CONV7 (ms) | CONV9 (ms) | CONV10 (ms) | CONV11 (ms) | CONV12 (ms) | CONV13 (ms) | Total Latency (ms) |
|-----------|------------|----------------------------------|------------------------|------------|------------|------------|------------|------------|------------|------------|------------|-------------|-------------|-------------|-----------------|
|           | 500MHz     | 168                              | 168                    | 1.515      | 3.348      | 1.92       | 1.44       | 0.96       | 1.42       | 13.0       | 14.597      | 27.071      | 29.81       | 28.473      | 9.184          |
|           | 200MHz     | 168                              | 168                    | 2.5        | 2.5        | 2.5        | 2.5        | 2.5        | 2.5        | 2.5        | 2.5        | 2.5         | 2.5         | 2.5         | 2.5             |

*Ratio = compablack work / our work for layer latency.  
bRatio = our work / compablack work for other.

TABLE IV: Layer by layer comparison for VGG-16.

| PE number | Clock rate | Peak throughput (clock rate * PE) | Real throughput (GOPS) | CONV1 (ms) | CONV2 (ms) | CONV3 (ms) | CONV4 (ms) | CONV5 (ms) | CONV7 (ms) | CONV9 (ms) | CONV10 (ms) | CONV11 (ms) | CONV12 (ms) | CONV13 (ms) | Total Latency (ms) |
|-----------|------------|----------------------------------|------------------------|------------|------------|------------|------------|------------|------------|------------|-------------|-------------|-------------|-------------|-----------------|
|           | 500MHz     | 168                              | 168                    | 1.515      | 3.348      | 1.92       | 1.44       | 0.96       | 1.42       | 13.0       | 14.597      | 27.071      | 29.81       | 28.473      | 9.184          |
|           | 200MHz     | 168                              | 168                    | 2.5        | 2.5        | 2.5        | 2.5        | 2.5        | 2.5        | 2.5        | 2.5         | 2.5         | 2.5         | 2.5         | 2.5             |

*Ratio = compablack work / our work for layer latency.  
bRatio = our work / compablack work for throughput.

2119-M-009-019, and Research of Excellence program 106-2633-E-009-001.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. NIPS, 2012, pp. 1097-1105.
[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," in Proc. CVPR, 2015, pp. 1-9.
[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in CoRR, 2014.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, 2016.
[5] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 221-231, 2013.
[6] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: integrated recognition, localization and detection using convolutional networks," in Proc. ICLR, 2014.
[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection nd semantic segmentation," in Proc. CVPR, 2014, pp. 580-587.
[8] T. He, W. Huang, Y. Qiao, and J. Yao, "Text attentional convolutional neural network for scene text detection," IEEE Transactions on Image Processing, vol. 25, pp. 529-2541, 2016.
[9] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in Proc. CVPR, 2015, pp. 5325-5334.
[10] Tomé, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro, "Deep convolutional neural networks for pedestrian detection," Signal Processing: Image Communication, 2016.
[11] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, et al., "Diannao: a small-footprint high-throughput accelerator for ubiquitous machine-learning," in Proc. ASPLOS, 2014, pp. 269-284.
[12] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, et al., "Dadiannao: a machine learning supercomputer," in MICRO, 2014, pp. 699-622.
[13] Z. Du, R. Fasthuber, T. Chen, P. Jenne, L. Li, T. Luo, et al., "ShiDianNao: shifting vision processing closer to the sensor," in ISCA, 2015, pp. 92-104.
[14] D. Liu, T. Chen, S. Liu, J. Zhou, S. Zhou, O. Teman, et al., "Pudianno: A polyvalent machine learning accelerator," in Proc. ASPLOS, 2015, pp. 369-381.
[15] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, et al., "A high-throughput neural network accelerator," in MICRO, 2015, pp. 24-32.
[16] Y.-H. Chen, T. Krishna, J. Emer, and V. Sze, "Eyerris: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," in ISSCC, 2016.
[17] Yu-Hsin Chen, Joel Emer, and Vivienne Sze, "Eyerris v2: A flexible for emerging deep neural networks on mobile devices," in JETCAS, 2019, pp. 292-308.
[18] F. Tu, S. Yin, P. Ouyang, S. Tang, L. Liu, and S. Wei, "Deep convolutional neural network architecture with reconfigurable computation patterns," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2017, pp. 2220-2233.
[19] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "Envision: A 0.26-to-1TOPS/W sub word parallel dynamic-voltage-accuracy-frequency scalable convolutional neural network processor in 28nm FDSOI," in ISSCC, 2017.
[20] Y.-J. Lin and T. S. Chang, "Data and hardware efficient design for convolutional neural network," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65, no. 1, pp. 198-208, 2018.
[21] L. Du et al., "A reconfigurable streaming deep convolutional neural network accelerator for internet of things," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65, no. 1, pp. 198-208, 2018.
[22] Y. Huan, J. Xu, L. Zheng, H. Tenhunen, and Z. Zou, "A 3D tiled low power accelerator for convolutional neural network," in ISCAS, 2018.
[23] D. Im et al., "DT-CNN: dilated and transposed convolution neural network for real-time image segmentation on mobile devices," in ISCAS, 2019.
[24] S. Liu et al., "Cambricon: An instruction set architecture for neural networks," in Proc. ISCA, 2016.
[25] S. Zhang et al., "Cambricon-X: An accelerator for sparse neural networks," in MICRO, 2016.
[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. NIPS, 2012, pp. 1097-1105.
[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. CVPR, 2014, pp. 580-587.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, 2016.
[29] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, pp. 221-231, 2013.
[30] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: integrated recognition, localization and detection using convolutional networks," in Proc. ICLR, 2014.
[31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection nd semantic segmentation," in Proc. CVPR, 2014, pp. 580-587.
[32] T. He, W. Huang, Y. Qiao, and J. Yao, "Text attentional convolutional neural network for scene text detection," IEEE Transactions on Image Processing, vol. 25, pp. 529-2541, 2016.
[33] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in Proc. CVPR, 2015, pp. 5325-5334.
[34] Tomé, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro, "Deep convolutional neural networks for pedestrian detection," Signal Processing: Image Communication, 2016.
deep neural network with pruning, trained quantization and huffman coding,” in ICLR, 2016.

[32] K. T. Malladi, B. C. Lee, F. A. Nothaft, C. Kozyrakis, K. Periyathamb, and M. Horowitz, “Towards energy-proportional datacenter memory with mobile dram,” in Proc. ISCA, 2012, pp. 37–48.

[33] H. Kwon, A. Samajdar, and T. Krishna, “MAERI: Enabling flexible dataflow mapping over DNN accelerators via reconfigurable interconnects,” in Proc ASPLOS, 2018, pp. 461-475.

[34] Y. Choi, D. Bae, J. Sim, S. Choi, M. Kim, and L. Kim, "Energy-efficient design of processing element for convolutional neural network,” in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 64, no. 11, pp. 1332-1336, 2017.

[35] J. Jo, S. Kim, and I. Park, "Energy-efficient convolution architecture based on rescheduled dataflow,” in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65, no. 12, pp. 4196-4207, 2018.

[36] A. Ardakani, C. Condo, M. Ahmadi, and W. J. Gross, "An architecture to accelerate convolution in deep neural networks,” in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 65(4), pp. 1349-1362, 2017.

[37] A. Biswas, and A.P. Chandrakasan, "CONV-SRAM: An Energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks." in IEEE Journal of Solid-State Circuits, vol. 54, no. 1, pp. 217-230, 2018.

[38] J. Wang, J. Lin, and Z. Wang, "Efficient hardware architectures for deep convolutional neural network." in IEEE Transactions on Circuits and Systems I: Regular Papers vol. 65, no. 6, pp. 1941-1953, 2017.

[39] A. Canziani, E. Culurciello, and A. Paszke, "An analysis of deep neural network models for practical applications." [Online]. Available: https://arxiv.org/abs/1605.07678

Kuo-Wei Chang is currently pursuing the Ph.D. degree with the National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C., and focus on VLSI design and hardware implementation for deep learning accelerators.

Tian-Sheuan Chang (S’93–M’06–SM’07) received the B.S., M.S., and Ph.D. degrees in electronic engineering from National Chiao-Tung University (NCTU), Hsinchu, Taiwan, in 1993, 1995, and 1999, respectively. From 2000 to 2004, he was a Deputy Manager with Global Unichip Corporation, Hsinchu, Taiwan. In 2004, he joined the Department of Electronics Engineering, NCTU, where he is currently a Professor. In 2009, he was a visiting scholar in IMEC, Belgium. His current research interests include system-on-a-chip design, VLSI signal processing, and computer architecture. Dr. Chang has received the Excellent Young Electrical Engineer from Chinese Institute of Electrical Engineering in 2007, and the Outstanding Young Scholar from Taiwan IC Design Society in 2010. He has been actively involved in many international conferences as an organizing committee or technical program committee member.