

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

Hyperdrive: A Multi-Chip Systolically Scalable Binary-Weight CNN Inference Engine

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Andri R., Cavigelli L., Rossi D., Benini L. (2019). Hyperdrive: A Multi-Chip Systolically Scalable Binary-Weight CNN Inference Engine. IEEE JOURNAL OF EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS, 9(2), 309-322 [10.1109/JETCAS.2019.2905654].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/703308> since: 2019-10-23

*Published:*

DOI: <http://doi.org/10.1109/JETCAS.2019.2905654>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the post peer-review accepted manuscript of:

R. Andri, L. Cavigelli, D. Rossi and L. Benini, "Hyperdrive: A Multi-Chip Systolically Scalable Binary-Weight CNN Inference Engine", in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 9, no. 2, pp. 309-322, June 2019. doi: 10.1109/JETCAS.2019.2905654

The published version is available online at: <https://doi.org/10.1109/JETCAS.2019.2905654>

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works

# Hyperdrive: A Multi-Chip Systolically Scalable Binary-Weight CNN Inference Engine

Renzo Andri\*, Lukas Cavigelli\*, Davide Rossi†, Luca Benini\*†

\*Integrated Systems Laboratory, ETH Zurich, Zurich, Switzerland †DEI, University of Bologna, Bologna, Italy

**Abstract**—Deep neural networks have achieved impressive results in computer vision and machine learning. Unfortunately, state-of-the-art networks are extremely compute and memory intensive which makes them unsuitable for mW-devices such as IoT end-nodes. Aggressive quantization of these networks dramatically reduces the computation and memory footprint. Binary-weight neural networks (BWNs) follow this trend, pushing weight quantization to the limit. Hardware accelerators for BWNs presented up to now have focused on core efficiency, disregarding I/O bandwidth and system-level efficiency that are crucial for deployment of accelerators in ultra-low power devices. We present Hyperdrive: a BWN accelerator dramatically reducing the I/O bandwidth exploiting a novel binary-weight streaming approach, which can be used for arbitrarily sized convolutional neural network architecture and input resolution by exploiting the natural scalability of the compute units both at chip-level and system-level by arranging Hyperdrive chips systolically in a 2D mesh while processing the entire feature map together in parallel. Hyperdrive achieves 4.3 TOPs/s/W system-level efficiency (i.e., including I/Os)— $3.1\times$  higher than state-of-the-art BWN accelerators, even if its core uses resource-intensive FP16 arithmetic for increased robustness.

**Index Terms**—Hardware Accelerator, Neural Network Hardware, Binary-Weight Neural Networks, Internet of Things, Systolic Arrays, Application Specific Integrated Circuits

## I. INTRODUCTION

Over the last few years, deep neural networks (DNNs) have revolutionized computer vision and data analytics. Particularly in computer vision, they have become the leading approach for the majority of tasks with rapidly growing data set sizes and problem complexity, achieving beyond-human accuracy in tasks like image classification. What started with image recognition for handwritten digits has moved to data sets with millions of images and 1000s of classes [1, 2]. What used to be image recognition on small images [3, 4] has evolved to object segmentation and detection [5–9] in high-resolution frames—and the next step, video analysis, is already starting to gain traction [10–12]. Many applications from automated surveillance to personalized interactive advertising and augmented reality have real-time constraints, such that the required computation can only be run on powerful GPU servers and data center accelerators such as Google’s TPUs [13].

At the same time, we observe the trend towards “internet of things” (IoT), where connected sensor nodes are becoming ubiquitous in our lives in the form of fitness trackers, smart phones and surveillance cameras [14, 15]. This creates a data deluge that is never analyzed and raises privacy concerns when collected at a central site [16]. Gathering all this data is

largely unfeasible as the cost of communication is very high in terms of network infrastructure, but also reliability, latency and ultimately available energy in mobile devices [17]. The centralized analysis in the cloud also does not solve the compute problem, it merely shifts it around, and service providers might not be willing to carry the processing cost while customers do not want to share their privacy-sensitive data [18].

A viable approach to address these issues is edge computing—analyzing the vast amount of data close to the sensor and transmitting only condensed highly informative data [14, 19]. This information is often many orders of magnitude smaller in size, e.g., a class ID instead of an image, or even only an alert every few days instead of a continuous video stream. However, this implies that the data analysis has to fit within the power constraints of IoT nodes which are often small-form factor devices with batteries of a limited capacity, or even devices deployed using a set-and-forget strategy with on-board energy harvesting (solar, thermal, kinetic, ...) [20].

Recently, several methods to train neural networks to withstand extreme quantization have been proposed, yielding the notions of binary- and ternary-weight networks (BWNs, TWNs) and binarized neural networks (BNNs) [21–23]. BWNs and TWNs allow a massive reduction of the data volume to store the network and have been applied to recent and high-complexity networks with an almost negligible loss. In parallel, the VLSI research community has been developing specialized hardware architectures focusing on data re-use with limited resources and optimizing arithmetic precision, exploiting weight and feature map (FM) sparsity, and performing on-the-fly data compression to ultimately maximize energy efficiency [24, 25]. However, these implementations fall into one of two categories: 1) They stream the entire or even partial FMs into and out of the accelerator ending up in a regime where I/O energy is far in excess of the energy spent on computation, hitting an energy efficiency wall: the state-of-the-art accelerator presented in [26] has a core energy efficiency of 59 TOPs/s/W, but including I/O power it is limited to 1 TOPs/s/W; or 2) they assume to store the entire network’s weights and intermediate FMs on-chip. This severely constrains the DNN’s size that can be handled efficiently by a small low-cost IoT-end node class chip. It also prevents the analysis of high-resolution images, thus precluding many relevant applications such as object detection.

The main contributions of this work are:

- 1) A new and highly optimized yet flexible core architecture systolically scalable to high-resolution images to enable applications such as object detection.

- 2) A new computational model, which exploits the reduced size of the weights due to the binarization in BWNs. As the size of the weights becomes much smaller than the intermediate feature maps, Hyperdrive streams the weights instead of the intermediate feature maps. With this new method, Hyperdrive enables execution of state-of-the-art BWNs on tiny, power-constrained chips, while overcoming the I/O energy-induced efficiency wall.
- 3) An in-depth analysis of this architecture in terms of memory requirements, I/O bandwidth, and scalability including measurements of the chip implemented in GF 22 nm FDX technology, showing a  $1.8\times$  and  $3.1\times$  gain in energy efficiency in image classification and object detection, respectively, even though our core uses resource-intensive FP16 arithmetic for increased robustness.
- 4) We show that the system is systolically scalable to multiple chips with the elementary chip size fixed to a maximum area constraint arranged in a 2D mesh operating on tiles of the entire feature map. The extension is also implemented in GF 22 nm FDX technology and is evaluated on layout simulations, showing that even with the overhead of exchanging the border pixels, the I/O energy can be reduced up to  $5.3\times$  compared with state-of-the-art accelerators.

The remainder of the paper is organized as follows. Sec. II presents a review of the previous works more closely related to the architecture presented in this paper. Sec. III and Sec. IV introduce the Hyperdrive architecture and computational model, respectively, mainly focusing on its key innovation aspect: stationary feature-map and streaming binary-weights for reduced I/O bandwidth and improved system-level energy efficiency. Sec. V describes the extensions to the presented architecture enabling a systolic-scalable system composed of Hyperdrive chips. Sec. VI presents the results of the chip implemented in 22nm FDX technology, providing details about its characterization, benchmarking, and comparison with respect to the state-of-the-art of binary-weight CNN accelerators. Finally, Sec. VII closes the paper with some final remarks.

## II. RELATED WORK

### A. Software-Programmable Platforms

The availability of affordable computing power on GPUs and large data sets have sparked the deep learning revolution, starting when AlexNet obtained a landslide victory in the ILSVRC image recognition challenge in 2012 [27]. Since then we have seen optimized implementations [28–30] and algorithmic advances such as FFT-based and Winograd convolutions further raising the throughput [31, 32]. The availability of easy-to-use deep learning frameworks (TensorFlow, Torch, Caffe, ...) exploiting the power of GPUs transparently to the user has resulted in wide-spread use of GPU computing. With the growing market size, improved hardware has become available as well: Nvidia has introduced a product line of systems-on-chip for embedded applications where ARM cores have been co-integrated with small GPUs for a power range of 5-20 W and  $\approx 50$  GOP/s/W. Also, the GPU's architecture

has been optimized for DNN workload, introducing tensor cores and fast half-precision floating-point (FP16) support. The latest device, Nvidia's V100, achieves 112 TFLOPS at 250 W [33]—an energy efficiency of 448 GOP/s/W. Its best known competitor, the first version of Google's TPU [13], works with 8-bit arithmetic and achieves 92 TOP/s at 384 W (240 GOP/s/W). With these power budgets, however, they are many orders of magnitude away from the power budget of IoT. Furthermore, they cannot exploit the advantages of many recent techniques to co-design DNN models for efficient computation.

### B. Co-Design of DNN Models and Hardware

Over the last few years, several approaches adapting DNNs to reduce the computational demand have been presented. One main direction was the reduction of the number of operations and model size. Specifically, the introduction of sparsity provides an opportunity to skip some operations. By pruning the weights a high sparsity can be achieved particularly for the fully-connected layers found at the end of many networks and the ReLU activations in most DNN models injects sparsity into the FMs, which can be exploited [34, 35].

A different direction is the research into reduced precision computation. Standard fixed-point approaches work down to 10-16 bit number formats for many networks. It is possible to further reduce the precision to 8 bit with small accuracy losses ( $< 1\%$ ) when retraining the network to adapt to this quantization [36]. There are limitations to this: 1) for deeper networks higher accuracy losses (2-3% for GoogLeNet) remain, and 2) Typically, only the inputs to the convolutions are quantized in this format. Internal computations are performed at full precision, which implies that the internal precision is very high for large networks, e.g., for a  $3\times 3$  convolution layer with 512 input FMs, this adds 12 bits. Further approaches include non-linearly spaced quantization in the form of mini-floats [36], and power-of-two quantization levels replacing multiplications with bit-shift operations (i.e., INQ [21]).

Several efforts have taken the path to extreme quantization to binary (+1/-1) or ternary (+1/0/-1) quantization of the weights while computing the FMs using floats. This massively compresses the data volume of the weights and has even been shown to be applicable to deep networks with an accuracy loss of approximately 1.6% for ResNet-18 [21] and thus less than the fixed point-and-retrain strategies. The next extreme approach are (fully) binary neural networks (BNNs), where the weights and FMs are binarized [37]. While this approach is attractive for extreme resource constrained devices [19, 38], the associated accuracy loss of 16% on ResNet-18 is unacceptable for many applications.

### C. FPGA and ASIC Accelerators

Many hardware architectures targeting DNNs have been published over the last few years. The peak compute energy efficiency for fixed-point CNN accelerators with precision bigger than 8 bit can be found at around 50 GOP/s/W for FPGAs, 2 TOP/s/W in 65 nm and around 10 TOP/s/W projected to 28 nm [30, 39–41].

Many of the sparsity-based optimizations mentioned in Sec. II-B have been implemented in hardware accelerators [35, 42] and achieve an up to  $3\times$  higher core energy efficiency and raise the device-level energy efficiency by around 70% through data compression. The effect of training DNNs to become BWNs simplifies the computations significantly and has shown the biggest impact on core compute-only energy with an energy efficiency of 60 TOP/s/W in 65 nm [26].

State-of-the-art silicon prototypes such as QUEST [43] or UNPU [44] are exploiting such strong quantization and voltage scaling and have been able to measure such high energy efficiency with their devices. The UNPU reaches an energy efficiency of 50.6 TOP/s/W at a throughput of 184 Gop/s with 1-bit weights and 16-bit activations on 16 mm<sup>2</sup> of silicon in 65 nm technology. However, all the aforementioned implementations, either don't consider the necessary I/O energy for streaming the FMs in the computation of energy efficiency, or they assume that intermediate results can be entirely stored in the limited-size on-chip memory. This restricts these devices to run networks capable of solving only low-complexity image recognition tasks or re-introduces a system-level energy efficiency wall at around 1 TOP/s/W as soon as the feature maps need to be streamed off-chip [26].

QUEST [43] addresses this issue by 3D-stacking 96 MB of SRAM distributed across 8 dies using inductive coupling for die-to-die wireless communication. They apply 4-bit logarithmic quantization to both the weights and the feature maps, which results in an accuracy drop in excess of what BWNs with high-precision feature maps achieve. In this configuration, they obtain an energy efficiency of 594 GOp/s/W while achieving a throughput of 1.96 TOP/s at 3.3 mW on a 122 mm<sup>2</sup> die in 40 nm technology.

Hyperdrive not only exploits the advantages of reduced weight memory requirements and computational complexity, but fundamentally differs from previous BWN accelerators [26, 44, 45]. The main concepts can be summarized as: 1) Feature Maps are stored entirely on-chip, instead the weights are streamed to the chip (i.e., feature map stationary). Thanks to the binary nature of the weights the overall I/O demand is reduced dramatically. 2) Through its hierarchically systolic-scalable structure it allows to efficiently scale to any sized feature map and even with silicon area restriction it is still scalable by tailing on a 2D mesh of Hyperdrive chips.

### III. HYPERDRIVE ARCHITECTURE

Hyperdrive is a scalable and flexible binary-weight neural networks accelerator that can be parametrized to fit a wide range of networks targeting a variety of tasks from classification to object detection. Fig. 1 shows a block diagram of Hyperdrive, where  $M \times N$  indicate the spatial parallelism (i.e., size of the FM), while  $C$  the output channel parallelism. It is composed of the following components:

- *Feature Map Memory (FMM)*: Is a multi-banked memory storing input and output FMs.
- *Array of  $C \times M \times N$  Tile Processing Units (TPUs)*: A single Tile-PU is illustrated in Fig. 2. It contains 1) a half-precision float adder/subtractor to accumulate partial

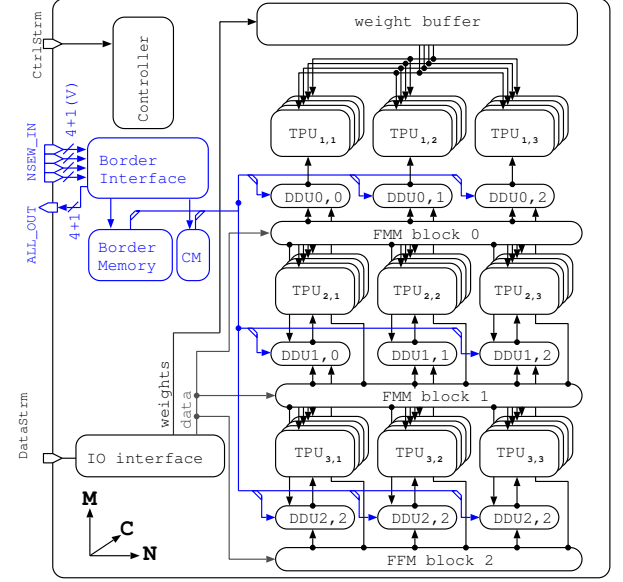


Fig. 1: System overview with  $C \times M \times N = 4 \times 3 \times 3$  tiles. Marked in blue are hardware block for the multi-chip systolic extension including the border interface which orchestrates any write and read to the border and corner memories and distributes it to the Data Distribution Units (DDUs). Furthermore, it sends and receives calculated pixels to and from the chip neighbors.

sums of the output pixels, bias and the bypass input FM (in case of residual blocks), 2) a half-precision multiplier for the FM-wise batch-normalization shared among the Tile-PU of the same tile, and 3) a ReLU activation unit. Each Tile-PU<sub>(c,x,y)</sub> is operating on the spatial tile  $(x,y)$  of the  $M \times N$  tiles and on the output channel  $c$  from  $C$ . Each Tile-PU is connected to its 8 spatial neighboring Tile-PU (i.e., directly adjacent Tile-PU) to quickly access neighboring pixels.

- *Weight Buffer (WBuf)*: Stores the weights of the current  $C$  output FMs.
- *Data Distribution Units (DDUs)*: Distributes the data from the memories to the corresponding Tile-PU units or manages zero-padding.
- *Border and Corner Memory BM, CM*: Storage for pixels which are part of neighboring chips.
- *Border Interface (BI/F)*: Sends and receive border pixels from/to neighboring chips and stores pixels into Border and Corner Memories.

The superior efficiency of Hyperdrive is achieved exploiting data re-use at different levels:

- *Output FM level*: The output FMs are tiled into blocks of  $C$  FMs which are calculated at the same time in the depth-wise parallel Tile-PU which allows to load the input FMs just once for  $C$
- *Spatial level*: The input FM is tiled into  $M \times N$  equally-sized image patches and calculated in parallel in the  $M \times N$  spatial processing units illustrated in Fig. 3. Weights are read once from off-chip memory only and used to calculate all  $M \times N$  partial sums for the corresponding tiles.
- *Weight re-use*: Weights are stored in the *weight buffer*,

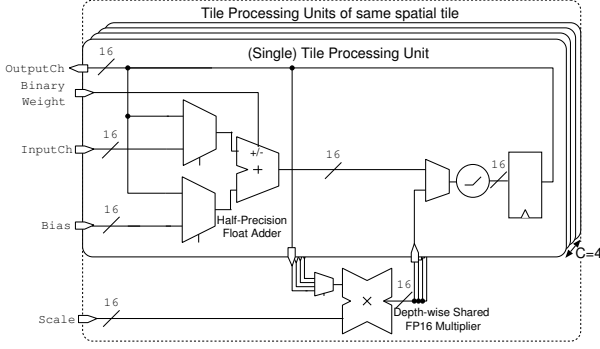


Fig. 2: *Tile Processing Units (TPUs) of same spatial tile* (Tile-PU $_{(\cdot, \cdot, y)}$ ): Every single Tile-PU (i.e., 4 shown in figure) provides a FP16 adder, accumulation register and ReLU activation unit. There is one time-shared FP16 multiplier per spatial tile and shared among the  $C = 4$  Tile-PU in the depth dimension, indicated by the dots. The FMs are calculated in a interleaved way for all  $C$  output dimensions. The (single-bit) binary weight is applied as the sign input for the FP16 adder.

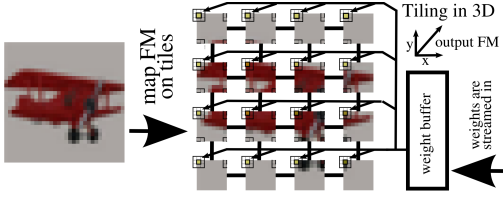


Fig. 3: The feature maps are tiled and processed in parallel Tile-PU's [46].

which is implemented as a latch-based standard cell memory for optimal energy efficiency [26].

- Border re-use: border pixels are transmitted only once to the corresponding neighbor chip and stored in its Border and Corner Memory instead of reading every time.

#### IV. COMPUTATIONAL MODEL

State-of-the-art CNNs like ResNet-34 impose high demands in computational complexity and memory for the large space of parameters and intermediate Feature Maps. However, for BWNs, streaming the weights rather than the FMs or both is particularly attractive due to the compression by  $16\times$  (i.e., from FP16).

CNNs are composed of several neural network layers, whereas the main building block are Convolution Layers which can be formulated as a mapping from the 3D input Feature Map space (i.e.,  $\text{FM}^{\text{in}}$ ) of  $n_{\text{in}}$  channels with  $h_{\text{in}} \times w_{\text{in}}$  sized spatial dimensions to the 3D output Feature Map space (i.e.,  $\text{FM}^{\text{out}}$ ) of  $n_{\text{out}} \times h_{\text{out}} \times w_{\text{out}}$  size and can be described as follows:

$$\mathbb{R}^{n_{\text{in}} \times h_{\text{in}} \times w_{\text{in}}} \xrightarrow{\text{CNN}} \mathbb{R}^{n_{\text{out}} \times h_{\text{out}} \times w_{\text{out}}}$$

$$\text{FM}^{\text{out}} \mapsto \text{FM}^{\text{in}} \text{ s.t.}$$

$$\text{FM}^{\text{out}}(c_{\text{out}}, \cdot, \cdot) = \beta_{c_{\text{out}}} + \alpha_{c_{\text{out}}} \sum_{c_{\text{in}} \in I_{n_i}} \text{FM}^{\text{in}}(c_{\text{in}}, \cdot, \cdot) * k_{c_{\text{out}}, c_{\text{in}}}(\cdot, \cdot)$$

Every single output channel  $c_{\text{out}}$  is calculated by convolving all input feature maps  $c_{\text{in}}$  with the corresponding filter kernel  $k_{c_{\text{out}}, c_{\text{in}}} \in \mathbb{R}^{h_k \times w_k}$ , scaled by the factor  $\alpha_{c_{\text{out}}}$  and accumulated to a bias term  $\beta_{c_{\text{out}}}$ . It should be noted here, that Batch normalization which are quite common after convolution layers, can be merged with biasing and scaling, as the coefficients stay constant after training.

#### A. Principles of Operation

The operations scheduling is summarized in Algorithm 1 and illustrated in Tbl. I for an implementation of the architecture featuring  $C \times M \times N = 16 \times 7 \times 7$  Tile-PU with  $8 \times 8$  sized spatial tiles  $\tilde{p}$  and for a  $3 \times 3$  convolution layer with  $16 \times 64$  FMs, whereas the output channels are tiled into blocks  $\tilde{c}_{\text{out}}$  of  $C = 16$  channels. After the entire input feature map is loaded into the FMM, the system starts inferring the network. The output FM-level and spatial parallelism is indicated in lines 2 and 3, whereas every Tile-PU is working on its assigned spatial tile  $\tilde{p}$  and output channel tile  $\tilde{c}$ .

Then in the inner loop, the contribution for all pixels from the corresponding tile and output channel are calculated. From the streaming approach, a logical approach would be to load the weights and apply it to the entire patch for every Tile-PU, unfortunately, the patches can be large, and this introduces frequent writes and reads to random-access memory (FMM), instead the weights streamed to the chip are stored in a weight buffer (Line 11) which can be implemented in a small memory (i.e., latch-based memory for low energy) and where the weights for the current  $C$  output channels (of all input channels) are stored. In this way, we avoid writing and re-fetching intermediate FM values.

The pixels are then calculated by iterating through all filter points (e.g., 9 in  $3 \times 3$  kernels) and input channels  $c_{\text{in}}$  (lines 7 and 8). On each cycle one binary weight per parallel feature map dimension  $\#\tilde{c}_{\text{out}}$  is loaded from the weight buffer (Line 14) and input Feature Map pixel per spatial tile ( $\#\tilde{p} = \#\{\text{Tile-PU}\} = M \cdot N$ ) are loaded from the FMM (Line 16). All the Tile-PU's access either their own FMM bank in case that the feature  $p + \Delta$  (for the filter tap  $\Delta$ , e.g.,  $(-1, -1)$  for the top-left weight of a  $3 \times 3$  filter) lies in the same tile  $\tilde{p}$  or from the corresponding FMM bank of the corresponding neighboring Tile-PU. All these accesses are aligned (e.g., all the Tile-PU's are reading the FMM bank of their corresponding top-left neighbor) and therefore no access conflicts occur. The weights are multiplied with the binary weights: this is implemented as a change of sign and then accumulated with the previous partial sum  $v$  (Line 17). When all contributions for all input channels and filter taps have been accumulated, a scaling factor (e.g., from batch normalization) is applied to it (Line 21), bypass is added (Line 22) and finally the channel bias is added (Line 23), before it is written back to the feature map memory (Line 24).

Bypass paths are common in several CNNs like ResNet-34 and are shown in Fig. 4. As will be explained in the next section, the bypass can be read, added to the partial sum and stored back to the same memory address avoiding additional memory for the bypass FM. Unfortunately, this does not work

in the same cycle, therefore adding the bias (Line 21) has been moved after the bypass (Line 20) and stalling can be avoided.

---

**Algorithm 1** Hyperdrive Execution-Flow
 

---

**Require:** All input feature maps in  $FMM^{in}$

**Require:** Weight Stream

```

1: for all  $M \times N$  pixel tiles  $\tilde{p}$  (in parallel HW units) do
2:   for all  $C$  output channel tiles  $\tilde{c}_{out}$  (in parallel HW units) do
3:     Tile-PU for output channel tile  $\tilde{c}_{out}$  and pixel tile  $\tilde{p}$ 
4:     def readFMfromMemory
5:       for all output channel  $c_{out}$  in tile  $\tilde{c}_{out}$  do
6:          $v = 0$ 
7:         for all pixel  $p = (y, x)$  in tile  $\tilde{p}$  do
8:           for all filter points  $\Delta = (\Delta y, \Delta x)$  with
              $\Delta y = -\lfloor \frac{h_k}{2} \rfloor, \dots, -1, 0, 1, \dots, \lfloor \frac{h_k}{2} \rfloor,$ 
              $\Delta x = -\lfloor \frac{w_k}{2} \rfloor, \dots, -1, 0, 1, \dots, \lfloor \frac{w_k}{2} \rfloor$  do
9:             for all input channel  $c_{in}$  do
10:              if  $w[c_{in}, c_{out}, \Delta] \notin WBuf$  then
11:                 $k_{c_{out}, c_{in}}(\Delta) = wghtStrm$ 
12:                 $WBuf[c_{in}, c_{out}, \Delta] = k_{c_{out}, c_{in}}(\Delta)$ 
13:              end if
14:               $w = WBuf[c_{in}, c_{out}, \Delta]$  (read of  $\#\tilde{c}_{out}$  bit)
15:              // Aligned read of  $FMM^{in}[p + \Delta, c_{in}]$  from
              // corresponding memory bank (either from
              // its own memory bank or the correspond-
              // ing neighbor's bank).
16:               $x = FMM^{in}[p + \Delta, c_{in}]$  (read of  $\#\tilde{p}$  words)
17:               $v = (v + x \cdot w) = \begin{cases} v + x & \text{if } w = 1 \\ v - x & \text{otherwise} \end{cases}$ 
18:            end for
19:          end for
20:        end for
21:        (opt)  $v \ast= bnorm(c_{out})$ 
22:        (opt)  $v \ast= FMM^{bypass}(c_{out}, p)$ 
23:        (opt)  $v \ast= bias(c_{out})$ 
24:         $FMM^{out}[c_{out}, p] = v$  (save in memory)
25:      end for
26:    end for
27:  end for

```

---

### B. CNN Mapping

The size of the on-chip memory for intermediate FM storage has to be selected depending on the convolution layer with the largest memory footprint of the network, hereinafter referred as *Worst-Case Layer* (WCL). Typically, the WCL is at the beginning of the network, since a common design pattern is to double the number of FMs after a few layers while performing at the same time a  $2 \times 2$  strided operation, thereby reducing the number of pixels by  $4 \times$  and the total FM volume by  $2 \times$ . To perform the computations layer-by-layer, avoiding usage of power hungry dual-port memories, we leverage a ping-pong buffer mechanism reading from one memory bank and writing the results to a different memory bank. Hence, for a generic CNN the amount of memory required by the WCL is:  $\max_{layers \text{ in CNN}} n_{in} h_{in} w_{in} + n_{out} h_{out} w_{out}$  words, since all input

and output FMs have to be stored to implement the described ping-pong buffering mechanism.

However, many networks have bypass paths, hence additional intermediate FMs have to be stored, as described in Fig. 4a for the potential WCLs of ResNet-34. This aspect has two implications:

- 1) In order to avoid additional memory (+50%), we perform an on-the-fly addition of the bypass path after the second  $3 \times 3$  convolution (i.e., the dashed rectangle is a single operation). This is done by performing a read-add-write operation on the target memory locations.
- 2) To avoid adding a stall cycle when reading and writing to the same memory area within the same cycle, the bias adding is moved after the bypass such that the following order is followed convolution, scale, bypass, bias, store back. In this way, the data can be read from memory address and stored back to the same address with one cycle latency.
- 3) The common transition pattern with the  $2 \times 2$ -strided convolution does not require additional memory. It temporarily needs three memory segments, but two of them are  $2 \times$  smaller and can fit into what has been a single memory segment before (M2 is split into two equal-size segments M2.1 and M2.2).

In the following section, the calculation of the WCL for ResNet-like networks with basic bypass blocks is discussed in detail and numbers are presented for ResNet-34, but does not limit the execution of networks with smaller WCL. To reduce off-chip data communication to a minimum, we will split the WCL in memory segments M1, M2, ... to indicate which data needs to be kept in on-chip memories at the same time. Hyperdrive always operates on a single convolutional layer at a time and is iterating several times over the entire input FM which therefore needs to be stored on-chip in memory section M1. The same is valid for the output FM which is calculated and stored in M2, respectively.

There are  $n_{out}$  output channels which have a  $h_{out} \times w_{out}$  sized output FM. These output FMs are calculated as sum of convolutions of every  $n_{in}$  input channel (with FMs size of  $h_{in} \times w_{in}$ ) on the  $h_k \times w_k$  sized filter kernels  $w_{k,n}$ .

For a normal convolution layer,

$$M = M1 + M2 = n_{in} \cdot h_{in} \cdot w_{in} + n_{out} \cdot h_{out} \cdot w_{out} \text{ [words]}$$

need to be stored, because the entire input FM is needed to calculate every single output FM.

In a next step, the special case with residual bypasses is evaluated like in ResNet [2] and similar residual networks. ResNet has two different types of residual blocks: the basic building block and the bottleneck building block. The basic building block is presented in Fig. 4.

Within the basic building block, there are two different cases, the first is  $n_{in} = n_{out}$  where there is no striding, thus also  $h_{in} = h_{out}$  and  $w_{in} = w_{out}$ . The input FM to the residual block will then be placed in the virtual memory section M1 and Hyperdrive computes the first  $3 \times 3$  convolution layer and writes the results into section M2, secondly Hyperdrive calculates the second convolutions layer reading from M2 and accumulating the output FM with the bypassed values in M1



TABLE I. Time schedule for a 16 input FM and 64 output FM  $3 \times 3$  convolution. Notation for filter weights:  $f_{\text{input FM, output FM}}^{\text{filter tap}(\Delta y, \Delta x)}$

cycle	1	2	...	16	17	...	...	144	145	...	288	...	9216	9217	...	36.8k	
weight input	$f_{1,(1-16)}^{-1,-1}$	$f_{2,\cdot}^{-1,-1}$	...	$f_{16,\cdot}^{-1,-1}$	$f_{1,\cdot}^{-1,0}$	...	...	$f_{16,\cdot}^{1,1}$	No I/O (loaded from weight buffer)					$f_{1,(17-32)}^{-1,-1}$	...	No I/O	
input FM	1	2	...	16	1	...	...	16	1	...	16	...	16	1	...	16	
filter tap pos.	-1,-1				-1,0			...	+1,+1	-1,-1	...	+1,+1	...	+1,+1	-1,-1	...	+1,+1
outp. pixel pos.	1,1								1,2				...	8,8	1,1	...	8,8
output FM	1-16 (in parallel)													17-32		...	49-64

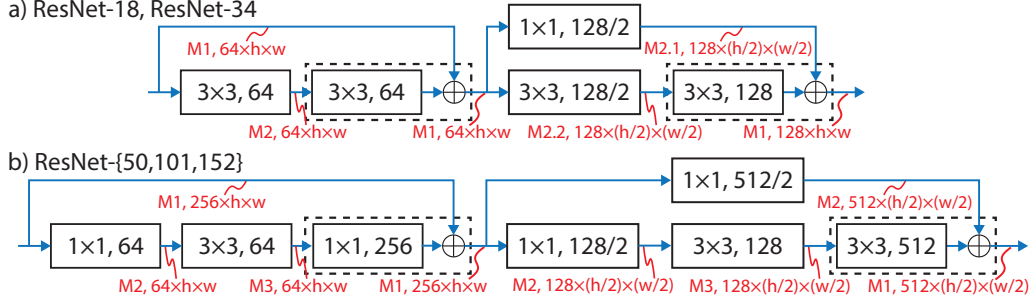


Fig. 4: Early block of layers of ResNet-34 and transition to next type of layer block. Activation and batch normalization layers are not indicated separately. Dashed rectangles imply on-the-fly addition to eliminate the need for additional memory

on-the-fly and writing them back to M1. A total amount of 401 kwords need to be stored.

$$\begin{aligned} M &= M1 + M2 = 2 \cdot M1 = 2n_{in} \cdot h_{in} \cdot w_{in} \\ M1 &= M2 = n_{in} \cdot h_{in} \cdot w_{in} \end{aligned}$$

$$M_{max} = 2n_{in} \cdot h_{in} \cdot w_{in} = 2 \cdot 64 \cdot 56 \cdot 56 = 401 \text{ kwords}$$

In case of down-sampling the number of output channels is doubled  $n_{out} = 2n_{in}$  and the image sizes are reduced by  $4 \times$  to  $h_{out} \times w_{out} = \frac{1}{2}h_{in} \times \frac{1}{2}w_{in}$ . Also, the bypass needs to be strided. He *et al.* suggest to either use the strided identity or to perform  $1 \times 1$  strided convolution, we will consider this case as it is more memory critical than with subsampling [2]. The input FM is read from M1 and the  $3 \times 3$  strided convolution is performed and saved in M2, then the  $1 \times 1$  strided convolution on the bypass is evaluated and saved in M3, finally the 2nd convolution layer is performed on the data in M2 and accumulated to the strided bypass in and to M3. It can be shown, that M2 and M3 are a quarter of the size of M1 and 301 kwords are needed for the three memory sections.

$$\begin{aligned} M &= M1 + M2 + M3 = 1.5 \cdot M1 \\ &= 1.5n_{in} \cdot h_{in} \cdot w_{in} \\ M1 &= n_{in} \cdot h_{in} \cdot w_{in} \\ M2 = M3 &= 2n_{in} \cdot 0.5 \cdot h_{in} \cdot 0.5 \cdot w_{in} = 0.5 \cdot M1 \end{aligned}$$

Due to the reduced size of the FM after every subsampling, just the first residual block need to be considered for dimensioning the memories. For ResNet-18 and ResNet-34, this translates to 401 kwords which are 6.4 Mbit with FP16.

Deeper residual networks (e.g., ResNet-50) are composed of the bottleneck building block (illustrated in Fig. 4b), to evaluate the WCL, there are two cases to consider: with and without subsampling. In the first case, the input FM is stored in M1 and needs to be stored for the entire bottleneck block. The

output FM for the first  $1 \times 1$  convolution layer is stored in M2 and is  $4 \times$  smaller due to the  $4 \times$  smaller number of channels, then the  $3 \times 3$  convolution layer calculates its features from M2 to M3 and the second  $1 \times 1$  convolution layer is calculated on-the-fly adding to the bypass FM.

$$\begin{aligned} M &= M1 + M2 + M3 = 1.5 \cdot M1 \\ &= 1.5n_{in} \cdot h_{in} \cdot w_{in} \end{aligned}$$

$$\begin{aligned} M1 &= n_{in} \cdot h_{in} \cdot w_{in} \\ M2 = M3 &= \frac{n_{in}}{4} \cdot h_{in} \cdot w_{in} = 0.5 \cdot M1 \end{aligned}$$

In total  $1.5 \times$  of the input FM size is needed to evaluate the bottleneck block without subsampling. In case with subsampling, already after the  $1 \times 1$  convolution, the bypass needs to be evaluated which is another  $1 \times 1$  convolution which we can map into M4 memory. Instead of writing the feature map for the  $3 \times 3$  convolution to M3, it can be written to M1, because this data is not needed any more. The 2<sup>nd</sup>  $1 \times 1$  convolution is then calculated on the fly from M1 and M4 back to M1.

$$\begin{aligned} M &= M1 + M2 + M4 = 1.675 \cdot M1 \\ &= \frac{13}{8}n_{in} \cdot h_{in} \cdot w_{in} = 1.2M \text{ words} \end{aligned}$$

$$\begin{aligned} M1 &= \max(n_{in} \cdot h_{in} \cdot w_{in}, \frac{2n_{in}}{4} \cdot \frac{h_{in}}{2} \cdot \frac{w_{in}}{2}) \\ &= n_{in} \cdot h_{in} \cdot w_{in} \end{aligned}$$

$$M2 = \frac{2n_{in}}{4} \cdot \frac{h_{in}}{2} \cdot \frac{w_{in}}{2} = 0.125 \cdot M1$$

$$M4 = 2n_{in} \cdot \frac{h_{in}}{2} \cdot \frac{w_{in}}{2} = 0.5 \cdot M1$$



TABLE II. Data Comparison for various typical networks with binary-weights and 16-bit FM's for single-chip implementation considering single-chip implementation (Top: Image Recognition, Bottom: Object Detection)

network	resolution	weights [bit]	all FM's [bit]	WC mem. [bit]
ResNet-18	224×224	11M	36M	6.4M
ResNet-34	224×224	21M	61M	6.4M
ResNet-50	224×224	21M	156M	21M
ResNet-152	224×224	55M	355M	21M
ResNet-34	2048×1024	21M	2.5G	267M
ResNet-152	2048×1024	55M	14.8G	878M

This leads to a WCL of 1.2 Mword or 19.2 Mbit (Conv2) for ResNet-50/-152/... independently of the depth which would be  $6.3 \text{ mm}^2$  of SRAM ( $0.3 \mu\text{m}^2/\text{bit}$  in GF 22nm FDX).

### C. Supported Neural Network Topologies

In the previous section, we have discussed the requirements to map the different ResNet-style networks onto Hyperdrive. For its implementation, we have parameterized the architecture to fit the feature maps of ResNet-34 on-chip. Nevertheless, Hyperdrive is neither restricted to these networks nor these applications—in fact, its scalability to multiple chips to process high-resolution images for object detection and image segmentation is a key feature of its architecture. For example, running the feature extraction for object detection using YOLOv2 [47] is supported by Hyperdrive. For the worst-case layer in terms of memory when processing  $448 \times 448$  pixel frames, we would need to be able to store 3.2 M words—scaling up the memory by  $2 \times$  over the ResNet-34 parameterization would be sufficient to even run it even on a single chip, and for higher resolutions the workload and memory for the feature maps could be split across multiple chips as described in Sec. V. Also, the Fire module of the size-optimized SqueezeNet [48] and SqueezeDet [5] topologies is supported by Hyperdrive. The grouped convolutions and shuffling operations present in MobileNetV2 [49] and ShuffleNet [50] can also be applied with the presented architecture. Also the not very common depth-wise separable convolutions present in some layers of MobileNetV2 can be computed using Hyperdrive, although not at maximum performance due to limited bandwidth of the on-chip SRAMs (no local re-use of the input feature map data possible).

The only limitation is that several networks feature a first convolution layer with an exceptionally large kernel size (e.g.,  $7 \times 7$  convolution for both ResNet and YOLOv2, but making up less than 2% of all operations). As Hyperdrive supports only  $1 \times 1$  and  $3 \times 3$  convolution layers, this first layer has to be computed off-chip before loading the data into Hyperdrive, or a small dedicated on-chip accelerator for the first layer could be included, which would perform these operations as the feature maps are streamed into the device. Networks optimized for compute effort, such as TinyYOLO [51] or MobileNetV2 [49], are often only composed of  $3 \times 3$  and  $1 \times 1$  convolution layers and do not have such a first filter with an exceptionally large kernel size.

## V. SCALABILITY TO MULTIPLE CHIPS

Even though, we could show that the architecture is in theory scalable to any sized networks, the WCL is setting a real-world limit to it. Already ResNets with bottleneck layer require 19.2 Mbit<sup>1</sup> to perform inference on small  $224 \times 224$  sized images and larger images (e.g., in typical object detection tasks) need 10s or 100s of Mbit. This clearly exceeds the area constraints of few Mbit in low-cost chip fabrication due to high production costs and diminished production yield. A very natural solution is to extend the systolic architecture to multiple chips, in this way the feature map is first tiled on an array of  $m \times n$  Hyperdrive chips and further tiled within each chip on their  $M \times N$  Tile Processing Units, such that  $M \cdot m \times N \cdot n$  tiles are operated in parallel.

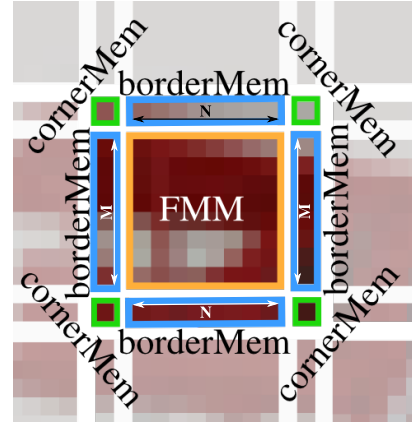


Fig. 5: Memory Allocation in the multi-chip setup with  $1 \times 1$  sized tiles for  $3 \times 3$  sized kernels. The  $M \times N$  “core” tiles and pixels are stored in the FMM and the pixels located and calculated in the chip neighbor are stored in Border and Corner Memory. The Border Memory stores these  $M \times$  or  $N \times$  pixels (i.e.,  $7 \times 16$  bit) which can be accessed in the same cycle.

Similarly to the single-chip setup, the Tile Processing Units need to access neighboring pixels, but in the multi-chip setup they might even lie on a different chip instead of just another tile memory. Three solutions are possible in this case, 1) the input feature maps of all chips are padded with the missing pixels, but this is not reasonable for deep networks as the padding increases steadily with the number of layers. 2) The border pixels are read from the neighboring chips when they are used, but this introduces high bandwidth requirement, as these pixels are needed several times or 3) the border pixels are sent once after they have been calculated to the neighboring chips and stored locally there. Hyperdrive implements option 3 which introduces two additional memories: A Border Memory BM and Corner Memory CM and have been added to the general architecture of Hyperdrive in Fig. 1.

Fig. 5 illustrates the locations of the pixels from a chip perspective and Fig. 6a shows the perspective of a single chip connected to its neighboring chips which are overall arranged in

<sup>1</sup>Note that the WCL for ResNet-like networks does not depend on depth, but on the size of the images (e.g.,  $224 \times 224$ ) and the building blocks (basic bypass in Fig. 4a or bottleneck in Fig. 4b). See also Tbl. II for a comparison of the WCLs.

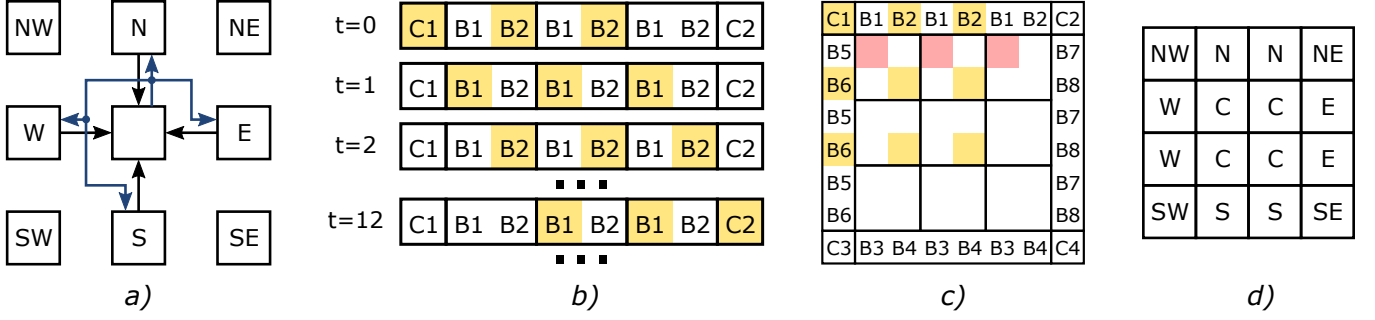


Fig. 6: Multi-chip Considerations: **a)** Intra-chip connection: 1 output interface and 4 inputs from/to 4 direct neighbors, **b)** Border Memory and Corner memory access with address block ( $c_{in} = 1, h_k = w_k = 3$ ) for every single cycle **c)** Access pattern in case of a corner access: two reads from Border Memory (top and left) and one read from Corner Memory **d)** Chip Types in a systolic chip setting (North West to South East and Center chip)

a systolic way. Pixels residing in the border of the neighboring chips are stored in the Border Memory and pixels residing in the corners of the diagonal neighboring chips are stored in the Corner Memory and are read from there in case border pixels are requested by the computational model.

#### A. Access Pattern and Storing Scheme of the Border Memories

Fig. 6c illustrates the pixels and their memory location which are read in case of a corner pixel and Fig. 6b for all cases of access top border pixels. When border pixels but not corner pixels have to be accessed, one pixel per corresponding border Tile-PU is read and stored into the same memory block. In case of a corner, actually  $M - 1$  and  $N - 1$  pixels from two border sides (i.e., one vertical and one horizontal) and one corner pixel. Therefore, the border memory is split into two physically separated memory blocks allowing to read from both sides without the need of two-port memories or introducing any latency or stalls. Furthermore, chips are assigned a location chip type, which indicates which part of the feature map the chip is working on. They have been named corresponding to cardinal orientation: corner chips (NW, NE, SW, SE), border chips (N, W, E, S) and Center like illustrated in Fig. 6d. All chips sharing the same orientation work identically and synchronized, thus the exact position does not matter.

#### B. Border and Corner Exchange

Whenever a border pixel (e.g., N border) has been calculated, it is sent to the corresponding neighbor (i.e., south neighbor) and a flag is set indicating that it is waiting the same kind of pixel from its opposite neighbor (i.e., north neighbor).

When a corner pixel (e.g., NW) is calculated, the pixel needs to be sent to all three neighboring chips in the corresponding direction (N, W, NW). As the number of these pixels is small and to keep the inter-chip wiring small, no additional diagonal interfaces are introduced, but these pixels are forwarded by the corresponding vertical neighbor (N) to the diagonal chip (NW). Additionally, there are for every corner 2 additional flags which are set in the Border Interface: one for the forwarding chip (sending, N) and the receiving chip (NW).

#### C. Border and Corner Memory

There are two different access patterns. If a corner pixel is accessed, the corner pixel,  $N - 1$  vertical pixels (left or right) and  $M - 1$  horizontal pixels (top or bottom) and one pixel need to be read from the corner memory, which is illustrated in Fig. 6c. In the other border cases, they are either  $N$  vertical pixels or  $M$  horizontal pixels (e.g., in Fig. 6b at  $t \in \{1, 2\}$ ). Therefore, the border memory can be seen as a horizontal or vertical extension to the FMM and  $N$  and  $M$  words can be read in a single cycle. As for the FMM, splitting the border memory into two physically separated memory blocks allows to read from both in the same cycle without introducing any additional latency. The memory needs to fit the overlapping border of the WCL whereas the width depends on the kernel width of the current and next layer. The overlapping rows or columns are  $\lfloor \frac{h_k}{2} \rfloor$  or  $\lfloor \frac{w_k}{2} \rfloor$  wide and can be determined directly from the WCL evaluation for FMM by dividing the spatial area and multiplying by the sum of all overlapping border rows or columns (which might differ for input and output FM). In case of ResNets with the basic building block (e.g., ResNet-34). The required memory for the left, right, top and bottom border (i.e.,  $M_{b,left}$ ,  $M_{b,right}$ ,  $M_{b,top}$ ,  $M_{b,bottom}$ ) can therefore be calculated as follows:

$$\begin{aligned}
 M_{border} &= M_{b,left} + M_{b,right} + M_{b,top} + M_{b,bottom} \\
 &= M \frac{2h_{in} + 2w_{in}}{h_{in} \cdot w_{in}} = M \frac{2 \cdot 56 + 2 \cdot 56}{56 \cdot 56} = 459 \text{ kbit} \\
 M_{b,left} &= M_{b,right} = 2 \left( n_{in} w_{in} \lfloor \frac{w_{k,l}}{2} \rfloor + n_{out} w_{out} \lfloor \frac{w_{k,l+1}}{2} \rfloor \right) \\
 M_{b,top} &= M_{b,bottom} = 2 \left( n_{in} h_{in} \lfloor \frac{h_{k,l}}{2} \rfloor + n_{out} h_{out} \lfloor \frac{h_{k,l+1}}{2} \rfloor \right)
 \end{aligned}$$

which is an increase of 7% of overall memory.

The Border Memory (as indicated in Fig. 1) is then implemented with 4 high-density single-port SRAMs with 1024 lines of  $7 \cdot 16 = 112$ .

The Corner Memory needs to store the diagonally overlapping pixels, which are  $\lfloor \frac{h_k}{2} \rfloor \cdot \lfloor \frac{w_k}{2} \rfloor$  sized patches. In contrary to the discussions regarding the FMM and BM, the Corner Memory does not profit from striding such that for ResNet typed networks the last layer has the highest memory

demand. Overall it can be dimensioned for ResNet-34 as  $(n_{in} + n_{out}) \cdot 4 \cdot \lfloor \frac{h_k}{2} \rfloor \cdot \lfloor \frac{w_k}{2} \rfloor = 2 \cdot 512 \cdot 4 \cdot 1 \cdot 1 \cdot 16 \text{ bit} = 64 \text{ kbit}$  which is another 1% increase of overall memory. This memory has been implemented with a single-port memory of 4096 of 16-bit words.

#### D. Interface Implementation

During the computation of border pixels, every border Tile-PU sends and receive the pixels to/from the respective Border Interfaces. The border interfaces, placed on the 4 sides (as illustrated in Fig. 6a) of the accelerator, are responsible for buffering and transmitting pixels from/to the neighboring chips, synchronizing execution of the Tile-PU as well. For vertical and horizontal borders there is one  $m \cdot C = 7 \cdot 16 = 112$  entries buffer. When the buffer is non-empty, the border interface sends these pixels in an interleaved way and split into blocks of 4 bits and 1 valid bit to the neighbors. Every chip itself has 4 in-coming serial interfaces from the directly adjacent neighbors (i.e., N, S, W, E). When data is received, it is de-serialized, recovered in its original 16-bit format and stored in the border/corner memories. The interfaces are also responsible for calculating the addresses of pixels received and transmitted from/to neighboring chips in the border memory. Fig. 2 shows in blue the extension needed for exchanging the borders between the chips with 1 out-going and 4 in-going intra-chip interfaces.

### VI. EXPERIMENTAL RESULTS

The number of tiles has been chosen to be  $M \times N = 7 \times 7$ , which allows for  $4 \times$  striding on  $112 \times 112$  sized input FMs (like in common ResNet-like networks), while keeping all the TPUs busy with at least one single spatial pixel during the entire network. We use the half-precision floating point (FP16) number format for the FMs as a conservative choice to ensure loss-less inference even for deeper networks [52, 53]. Fixed-point or other alternative formats [54] could be used to reduce the energy cost of the arithmetic operations. Fixed-point arithmetic units featuring a smaller number of bits (e.g., 8) would linearly impact the size of the on-chip memory for the FMs. By using FP16, the final accuracy is determined by the selected network and the corresponding BWN training algorithm. A ResNet-18 trained on the ImageNet dataset can run on Hyperdrive with a 87.1% top-5 accuracy using the SBD-FQ training method [55] (full-precision top-5 accuracy: 89.2%).

The on-chip memory was sized to fit the WCL of ResNet-34 with 6.4 Mbit (400 kword) and is implemented with  $M \times 8 = 7 \times 8$  high-density single-port SRAMs with 1024 lines of  $N \cdot 16 = 7 \cdot 16 = 112$ -bit words, whereas the memories are assigned to the  $(M \times N)$  tiles. The output FM parallelism has been fixed to  $C = 16$ . The weight buffer has been implemented to fit up to 512 (max. #input FMs)  $h_k \times w_k = 3 \times 3$  kernels for  $16 \times$  depth-wise parallelism. If more input FMs are needed, they can be tiled to 512 blocks and partial output FM can be calculated and summed up on-the-fly using the bypass mode. The frequently-accessed weight buffer has been implemented as a latch-based standard cell memory (SCM) composed of

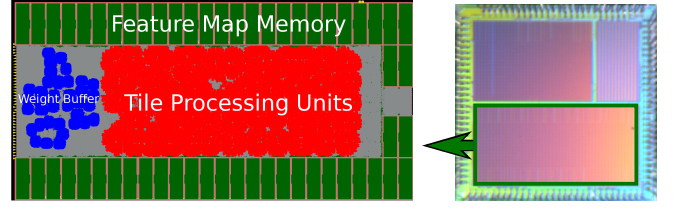


Fig. 7: Floorplan with Weight Buffer, Feature Map Memory and Tile Processing Units (left) and photograph of the taped-out multi-project chip Poseidon<sup>1</sup> with Hyperdrive on the bottom side.

$5 \times 8$  blocks of 128 rows of 16-bit words, reducing the access energy to SRAM memories by  $43 \times$  [26]. It should be noted that even though the energy efficiency of SCMs are much better than SRAMs, they are also up to  $8 \times$  larger in area which limits this kind of memories to comparably small buffers (i.e., weight buffer), but not for the feature map memory.

#### A. Implementation Results

Hyperdrive was designed in GF 22 nm FDX technology using an 8 track low voltage threshold (LVT) standard cell library. This flavor of the technology allows to apply up to 1.8V of forward body biasing (FBB), increasing the operating frequency of the chip at the cost of higher leakage power. Synthesis was performed with Synopsys Design Compiler 2017.09, while place & route was performed with Cadence Innovus 17.11.

The chip has an effective core area of  $1.92 \text{ mm}^2$  ( $\approx 9.6 \text{ MGE}$ )<sup>2</sup>, where  $1.24 \text{ mm}^2$  are SRAM memories (6.4 Mbit),  $0.115 \text{ mm}^2$  are SCM memory (74 kbit) and  $0.32 \text{ mm}^2$  arithmetic units. Fig. 7 shows on the right side a photograph of the actual chip and on the left side Hyperdrive's floorplan.

Testing and characterization (frequency, power) of silicon prototypes were performed on the industry-standard ASIC tester Advantest SoC V93000 and core power are based on the real chip measurements. The I/O energy was determined on the basis of an LPDDR3 PHY implemented in 28 nm technology [30], estimated as 21 pJ/bit, as in context of our research no low-swing interface IP blocks were available. It should be noted that this has to be considered as quite optimistic bound for I/O energy in a low-cost chip (the LPDDR3 PHY is quite complex and expensive), hence pessimistic for the proposed architecture focusing on system-level energy efficiency and specifically I/O bandwidth reduction. If we use low-cost low-complexity full-swing I/O interfaces (used for the implementation of this prototype, and of the other state-of-the-art accelerator [26, 35, 44, 45]) would further magnify the system-level energy gain of Hyperdrive with respect to other architectures, but would probably give too much advantage to our solution with respect to industrialized, production-ready scenario where low-swing I/O interfaces would be used [13].

Fig. 10 provides an overview of the Hyperdrive's blocks power consumption at the operating voltage of 0.5 V and

<sup>1</sup>Hyperdrive was taped-out alongside of two different projects (Kerbin and Quentin) on the same die to share costs, details can be found on <http://asic.ethz.ch/2018/Poseidon.html>

<sup>2</sup>One 2-input NAND gate equivalents (GE) is  $0.199 \mu\text{m}^2$  in GF22.

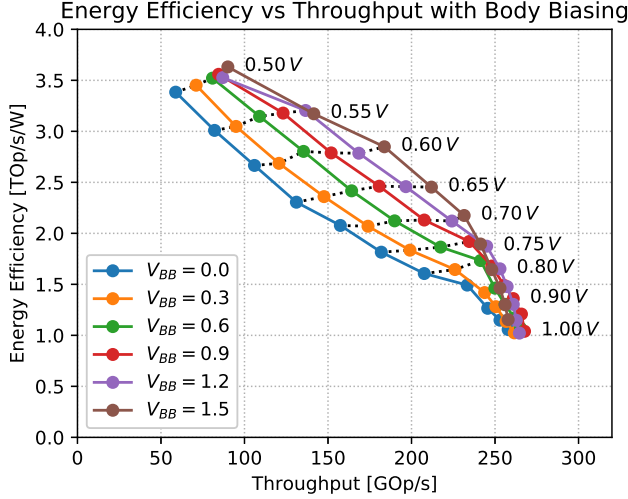


Fig. 8: Energy Efficiency vs. Throughput for different Body Bias Voltages including I/O for ResNet-34.

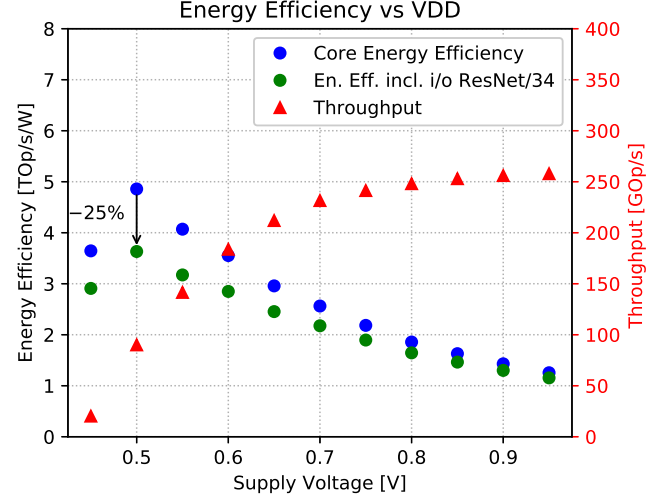


Fig. 9: Energy Efficiency and Throughput vs. supply voltages

58 MHz. The power consumption of memory arrays, memory periphery and logic were measured on the silicon prototype, available through the multi-power rails implementation strategy. On the other hand, the breakdown of the remaining standard cell logic power contributions is split into Tile-PUs, Weight Buffer and Others and has been estimated with post-layout simulations. It is interesting to note that a considerable amount of the power is consumed into the arithmetic units, while only a small overhead comes from memory accesses and I/Os, due to the efficient exploitation of Feature Map stationary (i.e., temporal locality) of the Hyperdrive architecture, explaining its superior system-level energy efficiency with respect to the other BWN accelerators in Tbl. V. The main features of the chip in other operating points is reported in Tbl. IV

In order to characterize the best energy point of the chip we swept the body bias of the system along the available range (i.e., from 0 V to 1.8 V), as shown in Fig. 8. It is interesting to note that both performance and energy efficiency increase together with body biasing, due to the favorable ratio between leakage power and dynamic power (4% at 0.5 V with no body biasing) and that even if the memory arrays are not body biased (i.e., leakage does not increase) the operating frequency increases significantly. This makes the operating points at 1.5 V FBB the most energy efficient ones for all performance targets. The best energy point occurs at 0.5 V VDD and 1.5 V FBB, featuring a throughput of 88 TOP/s and an energy efficiency of 3.6 TOPS/W running ResNet-34.

Fig. 9 shows the Energy Efficiency sweep vs. VDD. As mentioned before, the peak energy efficiency is achieved at 0.5V. Below this operating voltage, the relatively small operating frequency (i.e., 60 MHz) makes the leakage dominate, hence efficiency drops. It is interesting to note that, as opposed to other architectures implemented in scaled technologies, where the IO energy is dominating Tbl. V, in Hyperdrive the system level energy drops by only 25% when introducing the I/O energy into the analysis.

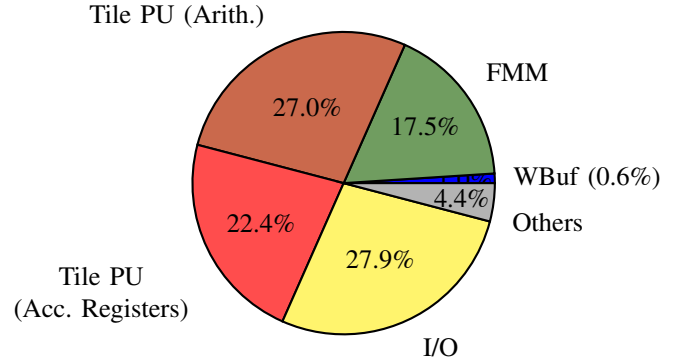


Fig. 10: Ratio of energy consumption at 0.5 V most energy-efficient corner.

TABLE III. Overview of Cycles, Throughput for ResNet-34

layer type	#cycles	#Op	#Op/cycle	#Op/s
conv	4.52 M	7.09 G	1568	
bnorm	59.90 k	2.94 M	49	
bias	59.90 k	2.94 M	49	
bypass	7.68 k	376.32 k	49	
total	4.65 M	7.10 G	1.53 k	431 G

TABLE IV. Overview of HYPERDRIVE (measured numbers)

Operating Point [V]	0.5	0.65	0.8
Op. Frequency [MHz]	57	135	158
Power [mW]	22	72	134
Throughput [Op/cycle]	1568	1568	1568
Throughput [GOp/s]	88	212	248
Core Energy Eff. [TOP/s/W]	4.9	3.0	1.9
Core Area [mm <sup>2</sup> ]	1.92	1.92	1.92
Memory [Mbit]	6.4	6.4	6.4



TABLE V. Comparison with State-of-the-Art BWN Accelerators (Top: Image Recognition, Bottom: Object Detection)

	Name	Techn.	DNN	Input Size	Precision Wghts/Acts	Core [V]	Eff. Th. [GOp/s]	Core E [mJ/im]	I/O E [mJ/im]	Total E [mJ/im]	En. Eff. [TOP/s/W]	Area [MGE]
Image Classification	YodaNN (layout) [26]	umc65	ResNet-34	224 <sup>2</sup>	Bin./Q12	1.20	490	0.9	3.6	4.5	1.6	1.3
	YodaNN (layout) [26]	umc65	ResNet-34	224 <sup>2</sup>	Bin./Q12	0.60	18	0.1	3.6	3.7	2.0	1.3
	Wang w/ 25 Mbit SRAM	SMIC130	ResNet-34	224 <sup>2</sup>	Bin./ENQ6	1.08	876	5.4	1.7	7.2	1.0	
	UNPU (chip)	65 nm	ResNet-34	224 <sup>2</sup>	Bin./Q16	0.77	346	2.3	3.6	6.0	1.2	11.1
	Hyperdrive (chip)	GF22	ResNet-34	224 <sup>2</sup>	Bin./FP16	0.50	88	1.4	0.5	1.9	3.6	9.0
	Hyperdrive (chip)	GF22	ResNet-34	224 <sup>2</sup>	Bin./FP16	1.00	263	6.5	0.5	7.0	1.0	9.0
	Wang w/ 25 Mbit SRAM	SMIC130	ShuffleNet	224 <sup>2</sup>	Bin./ENQ6	1.08	876	0.3	0.4	0.7	0.5	9.9
	UNPU (chip)	65 nm	ShuffleNet	224 <sup>2</sup>	Bin./Q16	0.77	346	0.1	1.0	1.1	0.3	11.1
	Hyperdrive (chip)	GF22	ShuffleNet	224 <sup>2</sup>	Bin./FP16	0.50	91	0.1	0.1	0.2	2.1	9.0
Object Detection	Wang w/ 25 Mbit SRAM	SMIC130	YOLOv3(COCO)	320 <sup>2</sup>	Bin./ENQ6	1.08	876	40.9	4.2	45.1	1.2	9.9
	UNPU (chip)	65 nm	YOLOv3	320 <sup>2</sup>	Bin./Q16	0.77	346	17.2	9.1	26.4	2.0	11.1
	Hyperdrive (chip)	GF22	YOLOv3	320 <sup>2</sup>	Bin./FP16	0.50	75	13.1	1.4	14.5	3.7	9.0
	Wang w/ 25 Mbit SRAM	SMIC130	ResNet-34	2k×1k	Bin./ENQ6			243.4	40.5	283.9	1.0	
	UNPU (chip) [44]	65 nm	ResNet-34	2k×1k	Bin./Q16	0.77	346	97.7	105.6	203.3	1.4	11.1
	Hyperdrive (10×5)	GF22	ResNet-34	2k×1k	Bin./FP16	0.50	4547	61.9	7.6	69.5	4.3	50×9.0
	Hyperdrive (20×10)	GF22	ResNet-152	2k×1k	Bin./FP16	0.50	18189	185.2	21.6	206.8	4.4	200×9.0
	Improvement over state-of-the-art for image classification (ResNet-34):								3.5×	1.8×	1.8×	
	Improvement over state-of-the-art for object detection: (ResNet-34):								5.3×	3.1×	3.1×	

### B. Benchmarking

The main evaluation of Hyperdrive has been performed on ResNet-34, whose network structure have been used in plenty of applications. This network features a good trade-off between depth and accuracy, i.e., ResNet-50 outperforms ResNet-34 by just 0.5% (Top-1) in terms of classification accuracy on the ImageNet dataset, but is roughly 50% more compute-intensive and the memory footprint is even 3.3× higher (see Sec. V).

The first and the last layer need to stay in full-precision to keep a satisfactory accuracy and are not implemented on Hyperdrive, but they contribute just 3% of the computation (226 MOp of 7.3 GOp) and can therefore also be evaluated on low-power compute platforms [56].

Tbl. III provides an overview of the number of operations, number of cycles and throughput while Hyperdrive is evaluating ResNet-34. In case of batch normalization, the throughput is reduced since just 49 multipliers are available and the normalization does take more cycles. In the layers where the bypass has to be added, Hyperdrive can also just calculate one output FM at a time, because the memory bandwidth is limited to 49 half-precision words. Fortunately, the non-convolution operations are comparably rare and a real throughput of 1.53 kOp/cycle or 221.9 GOp/s @ 0.65 V is achieved leading to a very high utilization ratio of 97.5% of the peak throughput. Tbl. VI provides an overview of the utilization (i.e., actual throughput normalized to theoretical peak throughput) for several networks. It can be seen that both ResNet-34 and ShuffleNet have very high utilization since the feature maps tile equally onto the Tile-PU. In the other case, where the intermediate feature maps are not sized by a multiple of  $M \times N$  (i.e.,  $7 \times 7$ ), the feature maps are padded with zeros and the last row and column of Tile-PU is idle during calculation of these zero pixels. Nevertheless, also in these cases, utilization is well above 80% (e.g., YOLOv3 [57] on a  $320 \times 320$  with 82.8%), which confirms the high flexibility of the proposed architecture with respect to different flavors of network topologies.

TABLE VI. Utilization of Hyperdrive

Network (Resolution)	#Op	#cycles	#Op/cycle	Utilization
Baseline (Peak Perf.)			1.57 k	100.0%
ResNet-34 (224 <sup>2</sup> )	7.10 G	4.65 M	1.53 k	97.5%
ShuffleNet (224 <sup>2</sup> )	140 M	90.3 k	1.55 k	98.8%
YOLOv3 (320 <sup>2</sup> )	53.1 G	33.9 M	1.30 k	82.8%

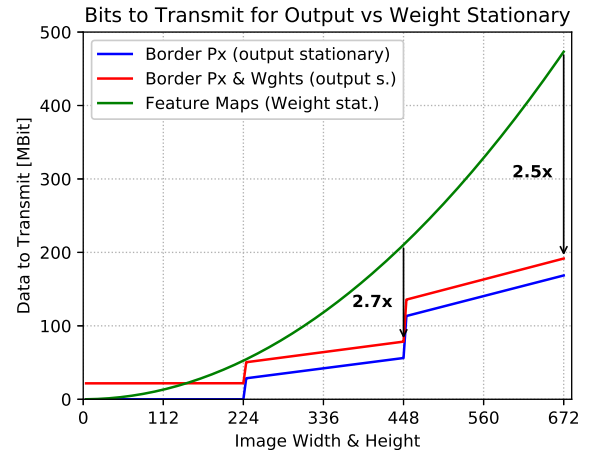


Fig. 11: Number of bits to be transmitted with the weight stationary approach compared to the output stationary approach adopted in the Hyperdrive architecture (including border exchange).

### C. I/O in Multi-Chip Setup

Having multiple-chips introduces implicitly more I/O as the border pixels have to be sent to the neighboring chips. To illustrate the relation between the feature map size to the amount of I/O, Fig. 11 compares the common weight stationary approach (green) to the feature map stationary approach of Hyperdrive (red). The evaluation is done with ResNet-34 with

the taped-out accelerator dimensioned to fit the WCL for  $3 \times 224 \times 224$  sized images. By scaling the spatial dimensions evenly, the amount of I/O stays constant for the weights of 21.6 Mbit until the maximum dimension of  $224 \times 224$  is reached. After that the FM is tiled onto several chips, starting with  $2 \times 2$ . This introduces the need exchange two entire rows and columns per output channel and layer to transmit and increases linearly with the FM size until the FM is not fitting anymore onto the  $2 \times 2$  chips, and tiling is done on  $3 \times 3$ , etc. In case of a systolic array of  $2 \times 2$  chips, the I/O can be reduced by up to  $2.7 \times$  and  $2.5 \times$  for a  $3 \times 3$  array while accounting for the border exchanges.

#### D. Comparison with State-of-the-Art

Tbl. V compares Hyperdrive with state-of-the-art binary weight CNN accelerators. The upper part of the table compares the SoA accelerators running image recognition applications (i.e., ResNet-34, VGG-16 and ShuffleNet on  $224 \times 224$  sized images), while the lower part compares key metrics coming from object detection applications with images available in autonomous driving data sets [5, 58] (i.e., ResNet-34 on  $2048 \times 1024$ , and YOLOv3 on  $320 \times 320$  images). At 0.65 V, Hyperdrive achieves a frame rate of 46.7 for ResNet-34, and, most important, the performance is independent of the image resolution thanks to the systolically scalable capabilities of the architecture.

While the totality of previous works is dominated by I/O energy, especially for spatially large feature maps, in Hyperdrive the I/O energy is only a small factor of the total energy (7% to 30%, depending on the application). Thanks to this feature, Hyperdrive outperforms other architectures by up to  $1.8 \times$  on image classification applications and up to  $3.1 \times$  in object detection applications, in terms of energy efficiency. More precisely, if we compare with the architecture presented in [45], Hyperdrive is  $3.1 \times$  more energy efficient, despite the number of bits used for the FMs in ENQ6 is only 6 [45], hence higher energy efficiency is achieved with much less aggressive reduction of HW precision. It should also be mentioned here, that previous work has estimated that for equi-precision results, highly discretized networks need to be just slightly larger (e.g., a ternary-weight (Q2) ResNet-18 is about 12% larger than a full-precision GoogLeNet while both achieve the same accuracy when trained with precision-aware algorithms [21]), whereas the core energy efficiency would improve significantly from stronger quantization and therefore Hyperdrive is expected to outperform the state-of-the-art even more than the  $3.1 \times$  factor reported here when using fixed-point representation and stronger quantization.

Furthermore, we compare our work with UNPU [44], which is the only silicon implementation adopting fixed-point arithmetic with adaptable precision (16, 8, 4, 2, 1) for the feature maps. We compare with the 16-bit mode, as this is the most similar with respect to accuracy. Our approach uses up to  $5.3 \times$  less energy for I/O and increases overall energy efficiency by up to  $3 \times$  since just the first input FM and the weights need to be streamed to the chip, but not the intermediate FMs. ShuffleNet is a challenging task for all the three accelerators

analyzed, as the feature maps are very deep, but spatially small. This implies a low compute intensity relative to the number of weights, which is an adverse pattern for Hyperdrive, and for most accelerators. On the other hand, grouping implies that for every group of output channels, just the subset of assigned input channels is filtered, which reduces the compute complexity while keeping the same feature map volume and is therefore an aspect in Hyperdrive's favor. Thus Hyperdrive still outperforms the state-of-the-art by  $4.2 \times$ .

The previous state-of-the-art accelerators are designed in less advanced technologies than Hyperdrive (GF 22nm compared to 65 nm and 130 nm), thus their core energy efficiency would be improved by using an advanced technology. Nevertheless, Hyperdrive's core energy efficiency is  $12.2 \times$  worse than YodaNN's and just  $1.6$  or  $3.7 \times$  better than UNPU and Wang *et al.* One of the reasons is that we use FP16 operators which are more robust than Q12 or ENQ6 in [26, 45] and were shown to work with the most challenging deep networks. Using floating-point feature maps directly impacts the energy for the accumulation operations as well as memory and register read/write operations. ENQ on the other side has been shown to introduce an accuracy drop of 1.6% already on CIFAR-100 [45], which is more than the difference between running ResNet-34 instead of ResNet-110 on CIFAR-10. It thus implies that a deeper network has to be computed to achieve a comparable accuracy. Furthermore, optimizations such as approximate adders and strong quantization have not been implemented, but can be combined with Hyperdrive's concepts, coupling core efficiency gains with the removal of the non-scalable I/O bottleneck. For instance, moving from FP16 to Q12 would lead to an energy efficiency boost that can be estimated to be around  $3 \times$  for the core, which would translate to a system efficiency boost of  $6.8 \times$  for high accuracy object detection with ResNet-34 features.

## VII. CONCLUSION

We have presented Hyperdrive: a systolically scalable hardware architecture for binary-weight neural networks, which dramatically minimizes the I/O energy consumption to achieve outstanding system-level energy efficiency. Hyperdrive achieves an energy efficiency of 4.3 TOP/s/W on object detection task which is more than  $3.1 \times$  better than prior state-of-the-art architectures, by exploiting a binary-weight streaming mechanism while keeping the entire FMs on-chip. Furthermore, while previous architectures were limited to some specific network sizes, Hyperdrive allows running networks not fitting on a single die, by arranging multiple chips in an on-board 2D systolic array, scaling-up the resolution of neural networks, hence enabling a new class of applications such as object detection on the edge of the IoT.

## ACKNOWLEDGEMENTS

This project was supported in part by the Swiss National Science Foundation under grant no. 162524 (MicroLearn: Micropower Deep Learning), armasuisse Science & Technology and the EU's H2020 program under grant no. 732631 (OPRECOMP).

## REFERENCES

- [1] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.
- [2] K. He *et al.*, “Deep Residual Learning for Image Recognition,” *Proc. IEEE CVPR*, pp. 770–778, 2015.
- [3] Y. LeCun *et al.*, “Convolutional Networks and Applications in Vision,” in *Proc. IEEE ISCAS*, 2010, pp. 253–256.
- [4] P. Sermanet *et al.*, “Convolutional Neural Networks Applied to House Numbers Digit Classification,” in *Proc. IEEE ICPR*, 2012, pp. 10–13.
- [5] B. Wu *et al.*, “SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving,” in *Proc. IEEE CVPRW*, 2017, pp. 446–454.
- [6] S. Ren *et al.*, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [7] J. Long *et al.*, “Fully Convolutional Networks for Semantic Segmentation,” in *Proc. IEEE CVPR*, 2015.
- [8] L. Cavigelli *et al.*, “Computationally efficient target classification in multispectral image data with Deep Neural Networks,” in *Proc. SPIE Security + Defence*, vol. 9997, 2016.
- [9] L. Cavigelli *et al.*, “CAS-CNN: A Deep Convolutional Neural Network for Image Compression Artifact Suppression,” 2016.
- [10] C. Feichtenhofer *et al.*, “Convolutional Two-Stream Network Fusion for Video Action Recognition,” in *Proc. IEEE CVPR*, 2016, pp. 1933–1941.
- [11] F. Scheidegger *et al.*, “Impact of temporal subsampling on accuracy and performance in practical video classification,” in *25th European Signal Processing Conference, EUSIPCO 2017*, vol. 2017-Janua, 2017.
- [12] L. Cavigelli *et al.*, “CBinfer: Change-Based Inference for Convolutional Neural Networks on Video Data,” in *Proc. ACM ICDSC*, 2017.
- [13] N. P. Jouppi *et al.*, “In-Datcenter Performance Analysis of a Tensor Processing Unit,” in *Proc. ACM ISCA*, 2017, pp. 1–17. [Online]. Available: <https://drive.google.com/file/d/0Bx4hafXDDq2EMzRNcy1vSUxtcEk/view>
- [14] F. Conti *et al.*, “An IoT Endpoint System-on-Chip for Secure and Energy-Efficient Near-Sensor Analytics,” *IEEE TCAS*, vol. 64, no. 9, pp. 2481–2494, 2017.
- [15] VentureBeat.com, “GreenWaves Technologies unveils Gap8 processor for AI at the edge,” 2018.
- [16] R. G. Baraniuk, “More Is Less: Signal Processing and the Data Deluge,” *Science*, vol. 331, no. 6018, pp. 717–719, 2011.
- [17] P. Schulz *et al.*, “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture,” *IEEE Comm. Mag.*, vol. 55, no. 2, pp. 70–78, 2017.
- [18] EE Times, “7 Ideas for AI Silicon from ISSCC – Google calls for hybrid edge/cloud collaboration,” 2018.
- [19] M. Rusci *et al.*, “Design Automation for Binarized Neural Networks: A Quantum Leap Opportunity?” in *Proc. IEEE ISCAS*, 2018.
- [20] A. S. Weddell *et al.*, “A Survey of Multi-Source Energy Harvesting Systems,” *Proc. ACM/IEEE DATE*, p. 4, 2013.
- [21] A. Zhou *et al.*, “Incremental Network Quantization: Towards Lossless CNNs with Low-Precision Weights,” in *Proc. ICLR*, 2017.
- [22] G. Venkatesh *et al.*, “Accelerating Deep Convolutional Networks using low-precision and sparsity,” in *Proc. IEEE ICASSP*, 2017, pp. 2861–2865.
- [23] M. Courbariaux *et al.*, “BinaryConnect: Training Deep Neural Networks with binary weights during propagations,” in *Adv. NIPS*, 11 2015, p. 9. [Online]. Available: <http://arxiv.org/abs/1511.00363>
- [24] V. Sze *et al.*, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
- [25] L. Cavigelli *et al.*, “Extended bit-plane compression for convolutional neural network accelerators,” *arXiv preprint arXiv:1810.03979*, 2018.
- [26] R. Andri *et al.*, “YodaNN: An Architecture for Ultra-Low Power Binary-Weight CNN Acceleration,” *IEEE TCAD*, 2017.
- [27] A. Krizhevsky *et al.*, “Imagenet Classification With Deep Convolutional Neural Networks,” in *Adv. NIPS*, 2012.
- [28] S. Chetlur *et al.*, “cuDNN: Efficient Primitives for Deep Learning,” in *arXiv:1410.0759*, 2014.
- [29] L. Cavigelli *et al.*, “Origami: A Convolutional Network Accelerator,” in *Proc. ACM GLSVLSI*. ACM Press, 2015, pp. 199–204.
- [30] L. Cavigelli *et al.*, “Origami: A 803-GOp/s/W Convolutional Network Accelerator,” *IEEE TCSVT*, vol. 27, no. 11, pp. 2461–2475, 2017.
- [31] N. Vasilache *et al.*, “Fast Convolutional Nets With fbfft: A GPU Performance Evaluation,” *arXiv:1412.7580*, 2014.
- [32] A. Lavin *et al.*, “Fast Algorithms for Convolutional Neural Networks,” in *Proc. IEEE CVPR*, 2016, pp. 4013–4021.
- [33] Nvidia Inc., “Nvidia Tesla V100 GPU Accelerator – Datasheet.”
- [34] T.-J. Yang *et al.*, “Designing Energy-Efficient Convolutional Neural Networks Using Energy-Aware Pruning,” in *Proc. IEEE CVPR*, 2017, pp. 6071–6079.
- [35] S. Han *et al.*, “EIE: Efficient Inference Engine on Compressed Deep Neural Network,” in *Proc. ACM/IEEE ISCA*, 2016, pp. 243–254.
- [36] P. Gysel *et al.*, “Hardware-Oriented Approximation of Convolutional Neural Networks,” *ICLR Workshops*, p. 8, 2016.
- [37] M. Courbariaux *et al.*, “Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1,” in *arXiv:1602.02830*, 2016.
- [38] A. Al Bahou *et al.*, “XNORBIN: A 95 TOP/s/W Hardware Accelerator for Binary Convolutional Neural Networks,” in *arXiv:1803.05849*, 2018.
- [39] Y.-H. Chen *et al.*, “Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks,” in *Proc. IEEE ISSCC*, 2016, pp. 262–263.
- [40] Z. Du *et al.*, “ShiDianNao: Shifting Vision Processing Closer to the Sensor,” in *Proc. ACM/IEEE ISCA*, 2015,



pp. 92–104.

- [41] F. Conti *et al.*, “A Ultra-Low-Energy Convolution Engine for Fast Brain-Inspired Vision in Multicore Clusters,” in *Proc. ACM/IEEE DATE*, 2015, pp. 683–688.
- [42] A. Aimar *et al.*, “NullHop: A Flexible Convolutional Neural Network Accelerator Based on Sparse Representations of Feature Maps,” *arXiv:1706.01406*, 2017.
- [43] K. Ueyoshi *et al.*, “Quest: A 7.49tops multi-purpose log-quantized dnn inference engine stacked on 96mb 3d sram using inductive-coupling technology in 40nm cmos,” in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 216–218.
- [44] J. Lee *et al.*, “Unpu: A 50.6tops/w unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision,” in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 218–220.
- [45] Y. Wang *et al.*, “An Energy-Efficient Architecture for Binary Weight Convolutional Neural Networks,” *IEEE TVLSI*, vol. 26, no. 2, pp. 280–293, 2017.
- [46] R. Andri *et al.*, “Hyperdrive: A Systolically Scalable Binary-Weight CNN Inference Engine for mW IoT End-Nodes,” in *2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 7 2018, pp. 509–515.
- [47] J. Redmon *et al.*, “YOLO9000: Better, Faster, Stronger,” in *Proc. CVPR*, 2017, pp. 187–213.
- [48] F. N. Iandola *et al.*, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5MB model size,” in *arXiv:1602.07360*, N. Navab *et al.*, Eds., vol. 9349, Cham, 2016.
- [49] M. Sandler *et al.*, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [50] X. Zhang *et al.*, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices,” in *arXiv:1707.01083*, 2017.
- [51] J. Redmon *et al.*, “You Only Look Once: Unified, Real-Time Object Detection,” 2016.
- [52] S. Migacz, “8-bit inference with tensorrt,” in *GPU Technology Conference*, 2017.
- [53] D. Das *et al.*, “Mixed precision training of convolutional neural networks using integer operations,” *arXiv preprint arXiv:1802.00930*, 2018.
- [54] G. Tagliavini *et al.*, “A transprecision floating-point platform for ultra-low power computing,” in *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2018, pp. 1051–1056.
- [55] Q. Hu *et al.*, “Training binary weight networks via semi-binary decomposition,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 637–653.
- [56] M. Gautschi *et al.*, “Near-Threshold RISC-V core with DSP extensions for scalable IoT endpoint devices,” *IEEE TVLSI*, vol. 25, no. 10, pp. 2700–2713, 2017.
- [57] J. Redmon *et al.*, “Yolov3: An incremental improvement,” *arXiv:1804.02767*, 2018.
- [58] M. Cordts *et al.*, “The Cityscapes Dataset for Semantic Urban Scene Understanding,” in *Proc. IEEE CVPR*, 2016,

pp. 3213–3223.



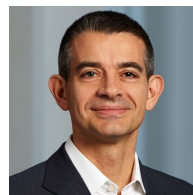
**Renzo Andri** received the M.Sc. degree in electrical engineering and information technology from ETH Zurich, Zurich, Switzerland, in 2015. He is currently pursuing a Ph.D. degree at the Integrated System Laboratory, ETH Zurich. His main research interests involve the design of low-power hardware accelerators for machine learning applications, and studying new algorithmic methods to further increase the energy-efficiency and therefore the usability of ML on energy-restricted devices.



**Lukas Cavigelli** received the B.Sc., M.Sc., and Ph.D. degree in electrical engineering and information technology from ETH Zürich, Zürich, Switzerland in 2012, 2014 and 2019, respectively. He has since been a postdoctoral researcher with ETH Zürich. His research interests include deep learning, computer vision, embedded systems, and low-power integrated circuit design. He has received the best paper award at the VLSI-SoC and the ICDSIC conferences in 2013 and 2017, and the best student paper award at the Security+Defense conference in 2016.



**Davide Rossi** received the Ph.D. from the University of Bologna, Italy, in 2012. He has been a post doc researcher in the Department of Electrical, Electronic and Information Engineering Guglielmo Marconi at the University of Bologna since 2015, where he currently holds an assistant professor position. His research interests focus on energy-efficient digital architectures in the domain of heterogeneous and reconfigurable multi- and many-core systems on a chip. This includes architectures, design implementation strategies, and run-time support to address performance, energy efficiency, and reliability issues of both high end embedded platforms and ultra-low-power computing platforms targeting the IoT domain. In these fields, he has published more than 80 papers in international peer-reviewed conferences and journals.



**Luca Benini** is the Chair of Digital Circuits and Systems at ETH Zurich and a Full Professor at the University of Bologna. He has served as Chief Architect for the Platform2012 in STMicroelectronics, Grenoble. Dr. Benini's research interests are in energy-efficient system and multi-core SoC design. He is also active in the area of energy-efficient smart sensors and sensor networks. He has published more than 1'000 papers in peer-reviewed international journals and conferences, four books and several book chapters. He is a Fellow of the ACM and of the IEEE and a member of the Academia Europaea.