

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

CUTIE: Beyond PetaOp/s/W Ternary DNN Inference Acceleration with Better-than-Binary Energy Efficiency

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

Published Version:

Moritz Scherer, Georg Rutishauser, Lukas Cavigelli, Luca Benini (2022). CUTIE: Beyond PetaOp/s/W Ternary DNN Inference Acceleration with Better-than-Binary Energy Efficiency. IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, 41(4), 1020-1033 [10.1109/tcad.2021.3075420].

Availability:

This version is available at: <https://hdl.handle.net/11585/905381> since: 2022-11-22

Published:

DOI: <http://doi.org/10.1109/tcad.2021.3075420>

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 final peer-reviewed accepted manuscript of:

M. Scherer, G. Rutishauser, L. Cavigelli and L. Benini

CUTIE: Beyond PetaOp/s/W Ternary DNN Inference Acceleration With Better-Than-Binary Energy Efficiency

In:

IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 4, pp. 1020-1033

The final published version is available online at:

<https://doi.org/10.1109/TCAD.2021.3075420>

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.

CUTIE: Beyond PetaOp/s/W Ternary DNN Inference Acceleration with Better-than-Binary Energy Efficiency

Moritz Scherer, Georg Rutishauser, Lukas Cavigelli, Luca Benini

Abstract—We present a 3.1 POp/s/W fully digital hardware accelerator for ternary neural networks. CUTIE, the Completely Unrolled Ternary Inference Engine, focuses on minimizing non-computational energy and switching activity so that dynamic power spent on storing (locally or globally) intermediate results is minimized. This is achieved by 1) a data path architecture completely unrolled in the feature map and filter dimensions to reduce switching activity by favoring silencing over iterative computation and maximizing data re-use, 2) targeting ternary neural networks which, in contrast to binary NNs, allow for sparse weights which reduce switching activity, and 3) introducing an optimized training method for higher sparsity of the filter weights, resulting in a further reduction of the switching activity. Compared with state-of-the-art accelerators, CUTIE achieves greater or equal accuracy while decreasing the overall core inference energy cost by a factor of $4.8 \times$ – $21 \times$.

Index Terms—Binary Neural Networks, Ternary Neural Networks, Hardware Accelerator, Deep Learning, Internet of Things, Application Specific Integrated Circuits

I. INTRODUCTION

Since the breakthrough success of AlexNet in the ILSVRC image recognition challenge in 2012 [1], Convolutional Neural Networks (CNNs) have become the standard algorithms for many machine learning applications, especially in the fields of audio and image processing. Supported by advances in both hardware technology and neural network architectures, dedicated Application-Specific Integrated Circuits (ASIC) hardware accelerators for inference have become increasingly commonplace, both in datacenter-scale applications as well as in consumer devices [2]. With the increasing demand to bring machine learning to Internet of Things (IoT) devices and sensor nodes at the very edge, the *de facto* default paradigm of cloud computing is being challenged. Neither are most data centers able to process the sheer amount of data generated by billions of sensor nodes nor can typical edge devices afford to send their raw sensor data to data centers for further processing, given their very limited power budget [3]. One solution to this dilemma is to increase the processing capabilities of each sensor node to enable it to only send extracted, highly compressed information over power-intensive

wireless communication interfaces or to act as an autonomous system.

However, the general-purpose microcontrollers typically employed in these IoT devices are ill-suited to the computationally intensive task of DNN inference, placing severe limitations on the achievable energy efficiency. While great strides in terms of energy efficiency have been made with specialized microcontrollers [4], some applications still require lower power consumption than what can be achieved with using 32-bit weights and activations in DNN inference. A popular approach to reducing the power consumption for neural network computations is the quantization of network parameters (weights) and intermediate results (activations). Quantized inference at a bit-width of 8 bits has been shown to offer equivalent statistical accuracy while allowing for significant savings in computation energy as well as reducing the requirements for working memory space, memory bandwidth, and storage by a factor of 4 compared to traditional 32-bit data formats [5], [6], [7], [8].

Pushing along the reduced bit-width direction, recently several methods to train neural networks with binary and ternary weights and activations have been proposed [9], [10], [11], [12], [13], [14], allowing for an even more significant decrease in the amount of memory required to run inference. In the context of neural networks, binary values refer to the set $\{-1, +1\}$ and ternary values refer to the set $\{-1, 0, 1\}$ [9], [15]. These methods have also been used to convert complex state-of-the-art models to their Binary Neural Network (BNN) or Ternary Neural Network (TNN) form. While this extreme quantization incurs sizeable losses in accuracy compared to the full-precision baselines, such networks have been shown to work well enough for many applications and the accuracy gap has been reducing quite rapidly over time [16], [17], [18].

Although quantization of networks does not affect the total number of operations for inference, it reduces the complexity of the required multipliers and adders, which leads to much lower energy consumption per operation. For binary networks, a multiplier can be implemented by a single XNOR-gate [19]. Further, the number of bit accesses per loaded value is minimized, which not only reduces the memory footprint but also the required wiring and memory access energy.

While Binary Neural Networks (BNNs) in particular are fairly well-suited to run on modern general-purpose computing platforms, to take full advantage of the potential energy savings enabled by aggressively quantized, specialized, digital, low-power hardware accelerators have been developed [20],

M. Scherer, G. Rutishauser, and L. Benini are with the Dept. of Information Technology and Electrical Engineering, ETH Zürich, Switzerland (e-mail: {scheremo, georg, benini}@iis.ee.ethz.ch).

L. Cavigelli is with Huawei Technologies, Zurich Research Center, Switzerland (e-mail: lukas.cavigelli@huawei.com).

L. Benini is also with the Dept. of Electrical, Electronic and Information Engineering, University of Bologna, Italy.

[19], [21], [22]. Concurrently to the research in digital neural network accelerators, analog accelerators that compute in-memory, as well as mixed-signal, have been explored [23], [24], [25]. While mixed-signal and in-memory designs hold the promise of higher energy efficiency than purely digital designs under nominal conditions, their higher sensitivity to process and noise variations, coupled with the necessity of interfacing with the digital world, are open challenges to achieve their full potential in energy efficiency [26].

Even though both analog and digital accelerators extract immense performance gains from the reduced complexity of each operation, there is still untapped potential to further increase efficiency. Most state-of-the-art binary accelerators use arrays of multipliers with large adder trees to perform the multiply-and-popcount operation [19], [21], [27], [25], which induces a large amount of switching activity in the adder tree, even when only a single input node is toggled. Adding to this, even state-of-the-art binary accelerators spend between 30% to 70% of their energy budget on data transfers from memories to compute units and vice-versa [25], [28]. This hurts efficiency considerably since time and energy spent on moving data from memories to compute units are not used to compute results. Taking these considerations into account, two major opportunities for optimization are to reduce switching activity in the compute units, especially the adder trees, and to reduce the amount of data transfer energy.

In this paper, we explore three key ideas to increase the core efficiency of digital low-bit-width neural network accelerator architectures: first, *unrolling of the data-path architecture with respect to the feature map and filter dimensions* leading to lower data transfer overheads and reduced switching activity compared to designs that implement iterative computations. Second, *focusing on Ternary Neural Networks (TNNs) instead of BNNs thereby capitalizing on sparsity to statistically decrease switching activity in unrolled compute units*. Third, *optimizing the quantization strategy of TNNs resulting in sparser networks that can be leveraged with an unrolled architecture*. We combine these ideas in CUTIE, the *Completely Unrolled Ternary Inference Engine*.

Our contributions to the growing field of energy-optimized aggressively quantized neural network accelerators are as follows:

- 1) We present the design and implementation of a novel accelerator architecture, which minimizes data movement energy spending by unrolling the compute architecture in the feature map and filter dimensions, demonstrating that non-computational energy spending can be reduced to less than 10% of the overall energy budget (Section V-C).
- 2) We demonstrate that by unrolling each compute unit completely and adjusting the quantization strategy, we directly exploit sparsity, minimizing switching activity in multipliers and adders, reducing the inference energy cost of ternarized networks by 36% with respect to their binarized variants (Section V-D).
- 3) We present analysis results, showing that the proposed architecture achieves up to 589 TOP/s/W in an IoT-suitable 22 nm technology and up to 3.1 POp/s/W in

an advanced 7 nm technology, outperforming the state-of-the-art in digital, as well as analog in-memory BNN accelerators, by a factor of $4.8\times$ in terms of energy per inference at iso-accuracy (Section V-G).

This paper is organized as follows: in Section II, previous work in the field of neural network hardware accelerators and aggressively quantized neural networks is discussed. In Section III, we introduce the proposed accelerator architecture. Section IV details the implementation of the architecture in the GlobalFoundries 22nm FDX and TSMC 7nm FF technologies. In Section V, the implementation results are presented and discussed, by comparing with previously published accelerators. Finally, Section VI concludes this paper, summarizing the results.

II. RELATED WORK

In the past few years, considerable research effort has been devoted to developing task-specific hardware architectures that enable both faster neural network inference as well as a reduction in energy per inference. A wide range of approaches to increase the energy-efficiency of accelerators have been studied, from architectural and device-level optimizations to sophisticated co-optimization of the neural network and the hardware platform.

A. Aggressively Quantized Neural Networks

On the algorithmic side, one of the main recent research directions has been quantization, i.e. representing model weights and intermediate activations in lower arithmetic precision. It has been known for some time that quantization of network weights to 5 bits and less is possible without a loss in accuracy in comparison to a 32-bit floating-point baseline model [5], [6], [7]. Further quantization of network weights to binary or ternary precision usually results in a small drop in accuracy, but precision is still adequate for many applications [12], [13], [29], [30]. Extending the approach of extreme quantization to intermediate activations, fully binarized and fully ternarized networks have been proposed [9], [15]. These types of networks perform very well on easier tasks such as 10-class classification on the well-established MNIST dataset [31], and efforts have been taken to improve their performance with novel training approaches [32], [33], [34]. Nevertheless, on more challenging tasks such as classification on the ILSVRC'12 dataset, they are still significantly less accurate than their full-precision counterparts [10], [35], [11], [17], [14], [36], [37]. Figure 1 depicts the accuracy gap between previously published, strongly quantized neural networks, their full-precision equivalents with identical architectures and the state-of-the-art full-precision networks on image classification tasks of increasing difficulty. On higher difficulty tasks, the gap between quantized networks and their full-precision equivalents grows larger. Furthermore, the gap between the full-precision architectures from which the quantized networks are derived and the overall state-of-the-art results reported in literature grows with task difficulty, indicating a prevalent focus in research activity on easier tasks and simple networks.

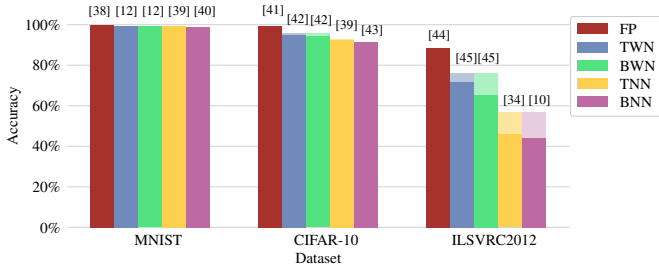


Fig. 1. Comparison of state-of-the-art accuracy of highly quantized neural networks of different precisions. **FP**: state-of-the-art in unquantized/full-precision neural networks, **BWN/TWN**: binary/ternary weight networks, **BNN/TNN**: fully binarized/ternarized neural networks. For the quantized network categories, the accuracy of the corresponding unquantized baseline networks is shown greyed out. As task difficulty is increased, a) the performance gap between the quantized networks and the full-precision baselines increases, and b) the gap between the unquantized baselines from which the quantized architectures are derived and the full-precision state-of-the-art widens.

Taking all of this into account, BNNs and TNNs provide a unique and interesting operating point for embedded devices, since they are by definition aggressively compressed, allowing for deep model architectures to be deployed to highly memory-constrained low-power embedded devices.

The core idea of binarization and ternarization of neural networks has been applied in numerous efforts, some of which also study the impact of the quantization strategy on the sparsity of ternary weight networks [13], [46], [47], [48]. While these previous efforts focus on the impact of the choice of quantization threshold and regularization, we evaluate the impact of quantization order, rather than threshold or regularization. Further, we study the effect of sparsity on the energy-efficiency of the proposed accelerator architecture.

B. DNN Hardware Accelerators

While the first hardware accelerators used for neural networks were general-purpose GPUs, there has been a steady trend pointing towards specialized hardware acceleration in machine learning in the past few years [49], [50], [51], [52]. Substantial research efforts have focused on exploring efficient architectures for networks using activations and weights with byte-precision or greater, [53], [54], [55], [22] different digital ASIC implementations for binary weight networks and BNNs have been proposed [20], [21], [56], [57], [58], [19]. Some works have tackled analog ASIC implementations of TNN accelerators, [23], [59], but very few digital implementations for TNN accelerators have been published [60], [61].

At the heart of every digital neural network accelerator lie the processing elements, which typically compute Multiply-Accumulate (MAC) operations. An important distinction between different architectures, besides the supported precision of their processing elements, lies in the way they schedule computations [49]. Most state-of-the-art architectures can be categorized into systolic arrays [53], [62], [56], [22], [23], which are flexible in how their processing elements are used, or output-stationary designs, which assign each output channel to one processing element [49], [21], [27]. Both approaches trade-off lower area for lower throughput and increased data

transfer energy by using iterative decomposition since partial results need to be stored and either weights or feature map data need to be reloaded. The alternative to iterative decomposition pursued in our approach, i.e. fully parallelizing the kernel-activation dot-products, is not only generally possible for convolutional neural networks, but also promises to be more efficient by increasing data-reuse and parallelism.

The state-of-the-art performance in terms of energy per operation for digital BNN and TNN accelerators is reported in Moons et al. [21] and Andri et al. [19], achieving peak efficiencies of around 230 TOP/s/W for 1-bit operations, as well as Knag et al. [27], reporting up to 617 TOP/s/W. The state-of-the-art for ternary neural networks is found in Jain et al. [23], achieving around 130 TOP/s/W for ternary operations.

In this work, we move beyond the state-of-the-art in highly quantized acceleration engines by implementing a completely unrolled data path. We show that by unrolling the data path, sparsity in TNNs is naturally exploited to reduce the required energy per operation without any additional overhead, unlike previous works [63], [64], [65], [66]. To capitalize on this effect, we introduce modifications to existing quantization strategies for TNNs, which are able to extract 53% more sparsity at iso-accuracy than by sparsity-unaware methods. Lastly, our work shows that ternary accelerators can significantly outperform binary accelerators both in terms of energy efficiency as well as statistical accuracy.

III. SYSTEM ARCHITECTURE

This section introduces the proposed system architecture. First, we present the data path and principle of operation and explain the levels of data re-use that the architecture enables, then we discuss considerations for lowering the overall power consumption. Finally, we present the supported functionality.

A. High-level Data Path

Figure 2 shows a high-level block diagram of the accelerator architecture. It is optimized for the energy-efficient layer-wise execution of neural networks. This is achieved first and foremost by a flat design hierarchy; each output feature map is computed channel-wise by dedicated compute units, called Output Channel Compute Unit (OCU). Each OCU is coupled with a private memory block for weight buffering, which minimizes addressing and multiplexing overheads for weight memory accesses, reducing the amount of energy spent on data transfers. The feature map storage buffers are shared between all OCUs to maximize the re-use of loaded activation data, which again aims to decrease the data transfer energy.

To exploit the high rate of data re-use possible with CNNs, the design uses a tile buffer, which produces tiles, i.e. square windows, of the input feature map in a sliding window manner. These windows are then broadcast to the pipelined OCUs.

An important aspect of aggressively quantized and mixed-precision accelerator design is choosing a proper compression scheme for its values. Since ternary values encode $\log_2(3) \approx 1.585$ bits per symbol, the most straight-forward compression approach would require 2 bits of memory per value, leaving

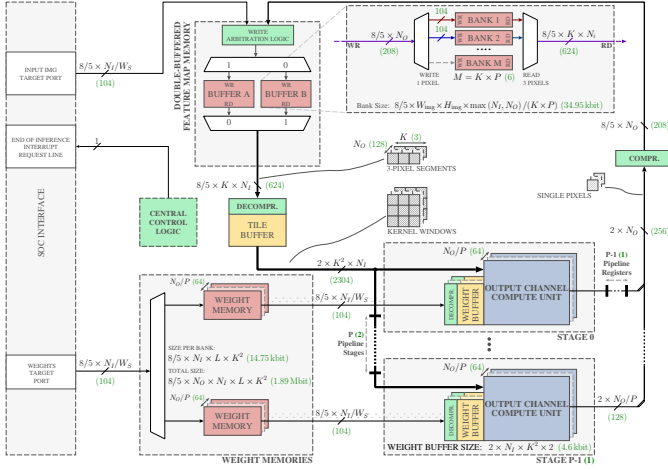


Fig. 2. Data-path schematic view of the accelerator core and its embedding into an SoC-level system. The diagram shows the unrolled compute architecture and encoding/decoding blocks, as well as the weight and feature map memories and tile buffer module. The dataflow of the accelerator is scheduled to first buffer full feature map windows in the tilebuffer and then compute the convolution result with pre-loaded weights in the compute units after which they are saved back to the feature map memory.

one of the four possible codewords unused. To reduce this overhead, values are stored 5 at a time, using 8 bits leading to 1.6 bits per symbol. The compression scheme used for this representation is taken from a recent work by Muller et al. [67]. To transition between the compressed representation and the standard 2's complement representation, compression and decompression banks are used with feature map and weight memories.

Figure 2 shows the pipeline arrangement of the OCUs. A key feature of the architecture is that an output channel computation is entirely performed on a single OCU. All OCUs need to receive input activation layers: the broadcast of input activations to OCUs is pipelined and the OCUs are grouped in stages. This pipeline fulfils multiple purposes: from a functional perspective, it allows to silence the input to clusters of compute units, which reduces switching activity during the execution of layers with fewer output channels than the maximum. Concerning the physical implementation of the design, pipelining helps to reduce fanout, which further reduces the overall power consumption of the design. It also reduces the propagation delay introduced by physical delays due to long wires.

B. Parametrization

The CUTIE architecture is parametrizable at compile time to support a large variety of design points. An overview of the design parameters is shown in Table I. Besides the parameters in Table I, the design's feature map memories and weight memories can be implemented using either Standard Cell Memories (SCMs) or SRAMs. CUTIE is designed to support arbitrary odd square kernel sizes K , pipeline depths P , input channel numbers N_I and output channel numbers N_O which directly dictate the dimensioning of the compute core, but also of the feature map memories and the tile buffer. The

TABLE I
DESIGN PARAMETERS OF CUTIE

Parameter	Description
N_I	Maximum number of channels of input feature map
N_O	Maximum number of channels of output feature map
K	Maximum kernel width and height
I_W	Maximum width of input feature map
I_H	Maximum height of input feature map
P	Maximum number of layers in the queue
P	Number of pipeline stages
W_S	Number of memory words per pixel

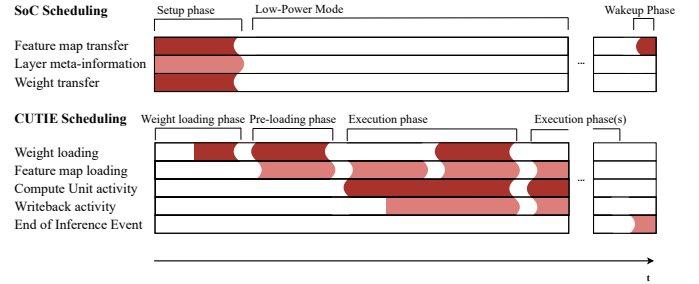


Fig. 3. Scheduling diagram of the accelerator core and SoC interface. The first two phases are needed to set up the first layer after reset, every other loading phase overlaps with an execution phase, which reduces the latency for scheduling a new layer to a single cycle. The host system can be put in a low-power mode while the accelerator core computes the network since all layer information is saved inside the core's memories.

OCU, as shown in Figure 4, consists of a compute core and a latch-based weight buffer that is designed to hold two kernels for the computation of one output channel, which amounts to $4 \times K^2 \times N_I$ bits. The feature map memories are designed to support the concurrent loading of K full pixels as well as the granular saving of $\frac{N_O}{P}$ ternary values. For these reasons, the word width of the feature map memories is chosen to be $\frac{N_O}{P}$ ternary values. To further allow for concurrent write and read accesses of up to K pixels, two feature map memories, each with $P \times K$ feature map memory banks, are implemented.

C. Principle of Operation

The accelerator core processes neural networks layer-wise. To enable layer-wise execution, networks have to be compiled and mapped to the core instruction set. The compilation process achieves two main goals: first, the networks' pooling layers are merged with the convolutional layers to produce fused convolutional layers. Second, the networks' convolutional layers' biases, batch normalization layers, and activation functions are combined to produce two thresholds that are used to ternarize intermediate results, similar to constant expression folding for BNNs [62]. After compilation, each layer consists of a convolutional layer with ternary weights, followed by optional pooling functions and finally, an activation function using two thresholds that ternarizes the result. To map the network to the accelerator, each layer's weights are stored consecutively in the weight memories, the thresholds are stored consecutively in the OCUs' Threshold FIFO and the meta-information like input width, stride, kernel size, padding, and so on are stored in the layer FIFO. All FIFOs, controllers

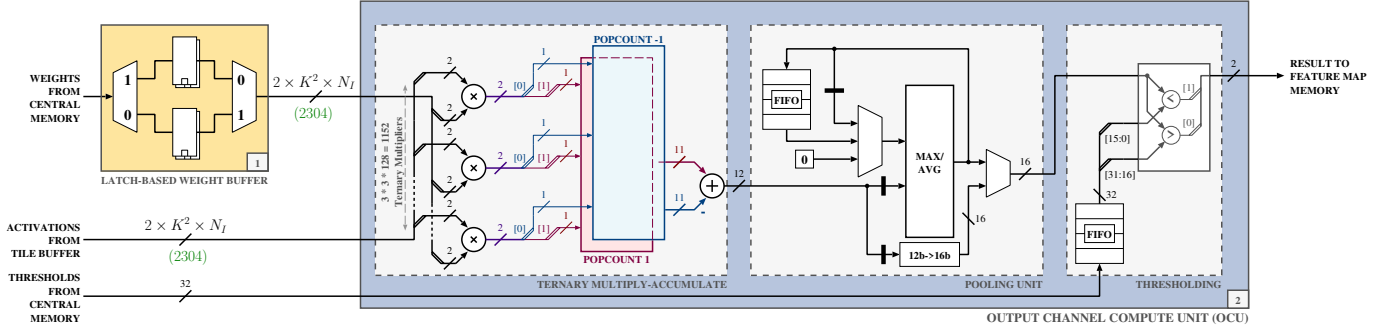


Fig. 4. Block diagram of the compute units for the design point $K = 3$, $N_I = N_O = 128$, showing the dual inner weight buffers (1), used for double buffering to avoid load stalling, the OCU (2), including the completely unrolled multiply/add tree, computing 1'152 multiply-accumulate operations in a single cycle, the pooling block, which enables max and average pooling and the thresholding module used to ternarize intermediate results. Notably, the multiplier and popcounts are fully combinational and not pipelined, which adds to the energy efficiency of the compute core.

and scheduling modules combined make up 2% of the total area.

The accelerator is designed to pre-buffer the weights for a full network during its setup phase and re-use the stored weights for multiple executions on different feature maps. Once at least one layer's meta-information is stored and the start signal is asserted, the accelerator's controllers schedule the execution of each layer in two phases; first, the weights for one layer are loaded into their respective buffers in the OCUs, then the layer is executed, i.e. every sliding window's result is computed and written back to the feature map memory. The loading of weights into the OCUs for the next layer and the computation of the current layer can overlap, leading to a single, fully concurrent execution phase after buffering the first set of weights, as shown in Figure 3. Once all layers have been executed, the end of inference signal is asserted, signalling to the host controller that the results are valid and the accelerator is ready for the next feature map input.

The module responsible for managing the loading and release of sliding windows is the tile buffer. The tile buffer consists of a memory array that stores K lines of pixel values implemented with standard cell latches. Feature maps are stored in a $(H \times W \times C)$ -aligned fashion in the feature map memory. To avoid load stalls and efficiently feed data to the compute core, up to K adjacent pixels at a time are read from the feature map memory. The load address is computed to always target the leftmost pixel of a window.

The scheduling algorithm for the release of the windows keeps track of the central pixel of the next-to-be scheduled window. This can be used to enable padding: for layers where padding is active, the scheduler starts the central pixel at the top left corner and zero-pads the undefined edges of the activation window. In case of no padding, the scheduler starts the central pixel to the lower-right of the padded starting position. For all but the first layer in a network, the weight loading and computation phases overlap such that the weights for the next layer are pre-loaded to eliminate additional loading latency.

The OCUs form the compute core of the accelerator. Figure 4 shows the block diagram of a single OCU. Each OCU contains two weight buffers, each of which is sized to hold

all the kernel weights of one layer. Having two buffers allows executing the current layer while also loading the next layer's weights. The actual computations are done in the ternary multipliers, each of which computes one product of a single weight and activation. While the input trits are encoded in the standard two's complement format, the result of this computation is encoded differently, i.e. the encoding is given by f :

$$f(x) = \begin{cases} 2'b10 & x = 1 \\ 2'b01 & x = -1 \\ 2'b00 & x = 0 \end{cases}$$

This encoding allows calculating the sum of all multiplications by counting the number of ones in the MSB and subtracting the number of ones in the LSB of all results, which is done in the popcount modules. The resulting value is stored as an intermediate result, either for further processing with the pooling module or as input for the threshold decoder. The threshold decoder compares the intermediate values against two programmable thresholds and returns a ternary value, depending on the result of the comparison. Notably, the OCU is almost exclusively combinational, requiring only one cycle of latency for non-pooling layers. Registers are only used to silence the pooling unit and in the pooling unit itself to keep a running record of the current pooling window. Since every compute unit computes one output channel pixel at a time, there are no partial sums that have to be written back.¹ However, to support pooling, each compute unit is equipped with a FIFO, a register, and an Add/Max ALU. In the case of max pooling, every newly computed value is compared to a previously computed maximum value for the window. In the case of average pooling, values are simply summed and the thresholds that are computed offline are scaled up accordingly. Figure 5 shows an example of the load & store schedule for pooling operations.

Low-power optimizations have been made on all levels of the design, spanning from the algorithmic design of the neural networks over the system architecture down to the choice of memory cells.

¹Which is a major difference from systolic arrays as well as output stationary designs!

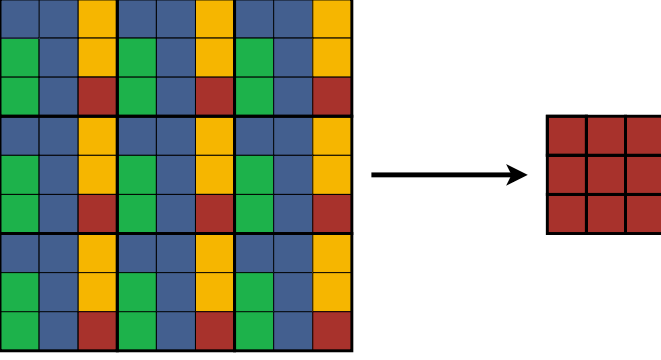


Fig. 5. Example of pooling buffer scheduling for 9×9 feature maps applying 3×3 pooling. The feature map is traversed left-to-right, top-to-bottom. Blue pixels are stored in the pooling unit's register, yellow pixels are stored in the pooling unit's FIFO for later use and green pixels are loaded from the pooling unit's FIFO and compared to the current value. Best viewed in color.

```

1 for w in range(featuremap_width):
2   for h in range(featuremap_height):
3     for co in range(output_channels):
4       for ci in range(input_channels):
5         for kw in range(kernel_width):
6           for kh in range(kernel_height):
7             out_fm[w][h][co] += in_fm[w+kw][h+kh][ci]
8             * kernel[kw][kh][ci][co]

```

Listing 1. Loop unrolling of convolutional layers implemented in the CUTIE architecture. The highlighted lines 3-8 are computed in parallel in a single shot, in combinational logic. Each OCU computes one output pixel channel value, i.e. each OCU computes one instance of the third loop.

Unlike most state-of-the-art architectures which use either systolic arrays or output-stationary scheduling approaches with iterative decomposition [53], [62], [56], [22], [23], [49], [21], [27], the CUTIE architecture unrolls the compute architecture fully with respect to weight buffering and output pixel computation, such that no storing of partial results is necessary; each output channel value is computed in a single cycle, as shown in Listing 1. The proposed design loads each data item exactly once and reduces overheads in multiplexing by clock gating unused modules. This applies to both the system level, with pipeline stages of the compute core that can be silenced, as well as to the module level, where the pooling module can be clock gated. To reduce both leakage and access energy, the feature map and weight memories can be implemented with standard cell latches, which are clock-gated down to the level of individual words. Generally, all flip-flops and latches in the design are clock-gated to reduce power consumption due to clock activity.

D. Input Encoding

To run real-world networks on the accelerator, the integer-valued input data has to be encoded with ternary values. We designed a novel ternary thermometer encoding based on the binary thermometer encoding [68]. The binary thermometer encoding is an encoding function f , that maps an integer between 0 and M to a binary vector with M entries.

$$f: \mathcal{N}_M \rightarrow \mathcal{B}^M$$

$$x \mapsto f(x)$$

$$f(x)_i = \begin{cases} 1 & i < x \\ -1 & i \geq x \end{cases}$$

The ternary thermometer encoding is an encoding function g that maps an integer between 0 and $2M$ to a ternary vector of size M .

$$g: \mathcal{N}_{2M} \rightarrow \mathcal{B}^M$$

$$x \mapsto g(x)$$

$$g(x)_i = \text{sgn}(x - M) \cdot \frac{f(|x - M|)_i + 1}{2}$$

The ternary thermometer encoding makes use of the additional value in the ternary number set with respect to the set of binary numbers and can encode inputs that are twice the size for a binary vector of a given size. The introduction of 0s in the encoding scheme further helps to reduce toggling activity in the compute units, lowering the average energy cost per operation. As an example, for $M = 128$, and $x = 110$ the binary thermometer encoding produces $[1]^{110} [-1]^{18}$, whereas the ternary thermometer encoding produces $[-1]^{18} [0]^{110}$.

E. Exemplary Instantiations of CUTIE

The architecture of CUTIE is highly parametric. In the following, we present two practical embodiments of the general architecture, which we will then push to full implementation. The instantiations of the accelerator presented in this section can process convolutions with a kernel of size 3×3 or smaller, using a stride between (1,1) and (3,3) with independent striding for the width and height dimension. It further supports average pooling and maximum pooling. Both no padding and full zero-padding, i.e. padding value of size 1 on every edge of feature maps, are supported. Depending on the requirements of the application, the feature map memory size and weight memory size should be configured to store the largest expected feature map and network. For the sake of evaluating the architecture, we chose to implement one version that supports feature maps up to a size of 32×32 pixels for both the current input feature map and the output feature map using SCMs and another version supporting sizes up to 160×120 feature map pixels using SRAMs. The supported feature map memory size does not restrict the functionality, since feature maps that do not fit within the memory can be processed in tiles. Assuming the feature maps need to be transferred from and to an external DRAM memory which requires 20 pJ/Bit, several orders of magnitude more energy than accessing internal memory, the critical goal is to minimize the amount of data transferred from and to external memory. To achieve that, we propose to adopt the depth-first computing schedule described in [69].

To estimate the energy cost of processing the feature map in tiles and to compare the layer-first and depth-first strategies on CUTIE, we compute the number of processed tiles per layer, the number of tiles that need to be transferred over the chip's I/O and the number of weight kernels that need to be switched for both the depth-first as well as the layer-first strategies. We assume a network consisting of eight convolutional layers using 3×3 kernels and 128 input and output channels. Using these results and simulated energy

TABLE II

ESTIMATED ENERGY CONSUMPTION OF A NETWORK CONSISTING OF 8 CONVOLUTIONAL LAYER WITHOUT POOLING FOR TILED COMPUTATION OF LARGE FEATURE MAPS ON A GF 22 SCM IMPLEMENTATION INCLUDING I/O AND EXTERNAL DRAM

	Depth-first	Layer-first
32×32	7.3 μ J	7.3 μ J
Bit accesses from or to external memory	209 kB	209 kB
Feature map transfer energy	4.2 μ J	4.2 μ J
Weight memory transfer energy	0.3 μ J	0.3 μ J
Computational energy	2.8 μ J	2.8 μ J
64×64	277 μ J	1'069 μ J
Bits moved from or to external memory	12.6 MB	52.8 MB
Feature map transfer energy	252 μ J	1'057 μ J
Weight memory transfer energy	2.5 μ J	0.3 μ J
Computational energy	22.5 μ J	11.5 μ J
96×96	3'734.5 μ J	6'030.3 μ J
Bit accesses from or to external memory	179.3 MB	300.1 MB
Feature map transfer energy	3'586 μ J	6'002 μ J
Weight memory transfer energy	14.5 μ J	0.3 μ J
Computational energy	134 μ J	28 μ J

costs for computations and memory transfers, we compute the additional cost when processing large feature maps layer- and depth-wise. For large frames, the cost is clearly dominated by the external memory access energy. Table IV shows an exploration over different frame sizes starting from 32×32 for which no tiling is required and extending to 64×64 and 96×96 that require significant external memory transfer. We find that by minimizing the feature map movement, the depth-first strategy consumes significantly less than the layer-first strategy for practical cases.

While the CUTIE core is designed to be integrated with a host processor, one key idea to reduce system-level energy consumption realized in the architecture is the autonomous operation of the accelerator core. The control implementation allows the accelerator to compute a complete network without interaction with the host. In the presented version, the weight memories, the layer FIFO, and threshold FIFOs are designed to store up to eight full layers, which can be scheduled one after another without any further input. In general, the number of layers can be freely configured, at the cost of additional FIFO and weight memory.

Besides offering support for standard convolutional layers, the architecture can be used for depthwise convolutional layers by using weight kernels where each kernel is all zeros except for one channel. Further, it can be used for ternary dense layers with input size smaller or equal to $3 \times 3 \times 128 = 1'152$ and output size smaller or equal to 128 by mapping all dense layer matrix weights to the $3 \times 3 \times 128$ weight buffer of an OCU.

IV. IMPLEMENTATION

This section discusses the implementation of the CUTIE accelerator architecture. The results from physical layouts in a 22 nm technology, one using SCMs and another using SRAMs, and from synthesis in a 7 nm technology are presented and discussed.

A. Interface Design

The interface of the accelerator consists of a layer instruction queue and read/write interfaces to the feature map and weight memories. The interface is designed to allow integration into a System-on-Chip (SoC) design targeting near-sensor processing. In this context, a pre-processing module could be connected to a sensor interface, with a host processor only managing the initial setup and off-chip communication. This setup consists of writing the weights into their respective weight memories and pre-loading the layer instructions into the instruction queue. In the actual execution phase, i.e. once data is loaded continuously, the accelerator is designed to autonomously execute the layer instructions without needing any further input besides the input feature maps and return only a highly-compressed feature map or even final labels. The end of computation is signalled by a single-bit interrupt to the host.

B. Dimensioning

The CUTIE architecture is not architecturally constrained to support a certain number of input/output channels, i.e. it can be parameterized to support an arbitrary amount of channels. Since it can be synthesized with support for any number of channels and feature map sizes, the proposed implementation was designed to optimize the accuracy vs. energy efficiency trade-off for the CIFAR-10 dataset. To this end, the compute units were synthesized and routed for different channel numbers to evaluate the impact of channel number on the energy efficiency of individual compute units and by extension, the whole accelerator. The estimations were performed for 64, 128, 256, and 512 channels. To estimate the energy efficiency of the individual implementations, a post-layout power simulation was performed, using randomly generated activations and weights. This experiment was repeated and averaged over 300 cycles, i.e. 300 independently randomly generated weight tensors and feature maps were used. Further, post-synthesis simulation estimations for the energy cost of memory accesses, encoding & decoding, and the buffering of activations and weights were added. The estimations for the resulting accelerator-level energy efficiency are shown in Figure 6. Since these estimations were made using a post-layout power simulation of a single OCU, they take into account the wiring overheads introduced by following the completely unrolled compute architecture. One of the main drivers for lower efficiency in the designs with more channels is the decrease in layout density and an increase in wiring overheads. While energy efficiency per operation does not directly imply energy per inference, it is a strong indicator of system-level efficiency.

C. Implementation Metrics

The accelerator design was implemented with a full back-end flow in GlobalFoundries 22 nm FDX and synthesized in TSMC 7 nm technology. The first of two implementations based on GlobalFoundries 22 nm FDX was synthesized using SRAMs supplied with 0.8 V for feature map and weight

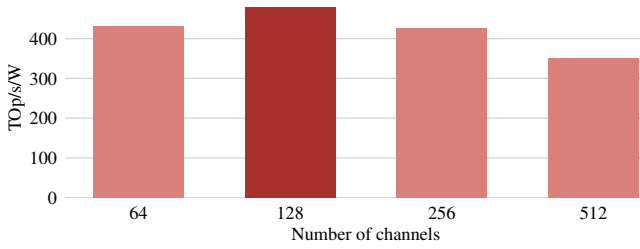


Fig. 6. Estimation of accelerator-level energy efficiency using data from the simulation of single OCUs, assuming SCM-based memories. Feature maps and weights were drawn from a uniform random distributions. There is a peak in energy efficiency at 128 channels before falling off for increasing channel numbers.

memories and 8 track standard cells operating at 0.65 V. The second of the GF 22 nm implementations uses SCM-based feature map and weight memories as well as 8 track standard cells for its logic cells, all supplied with 0.65 V. The TSMC 7 nm implementation similarly uses SCM-based memories to allow for voltage scaling. The post-synthesis timing reports show that the GF 22 nm implementations should be able to operate at up to 250 MHz. We chose to run both the SCM as well as the SRAM implementation at a very conservative frequency of 66 MHz. Since we did not run a full backend implementation of the 7 nm version, we chose to estimate the performance at the same clock frequency and voltage as the 22 nm versions. The total area required by the design is 7.5 mm^2 for both 22 nm implementations and approximately 1.2 mm^2 at a layout density of 0.75 for the 7 nm implementation. The reason for both GF 22 nm implementations requiring the same amount of area is due to the larger memories supported in the SRAM implementation, as explained in section III-E. A breakdown of the area usage in the SCM-based 22 nm implementation is shown in Figure 7.

For the GF 22 nm implementations, the sequential and memory cells take up around 80% of the overall design's area, while the clock buffers and inverters constitute only a very small amount of the total area. This characteristic is due to the choice of using latch-based buffers for a lot of the design and clocking the accelerator at a comparatively low frequency, while also extensively making use of clock-gating at every level of the design's hierarchy. Note that even though the area of the design is storage-dominated, power and energy are not, which is one of the key reasons for the extreme energy efficiency of CUTIE.

V. RESULTS AND DISCUSSION

This section discusses the evaluation results of the proposed accelerator design. First, we discuss the design and training of the network that is used to evaluate the accelerator's performance. Next, we discuss the general evaluation setup. Finally, we present the implementation and performance metrics and compare our design to previous work.

A. Quantized Network Training

The accelerator was evaluated using a binarized and a ternarized version of a neural network, using the binary

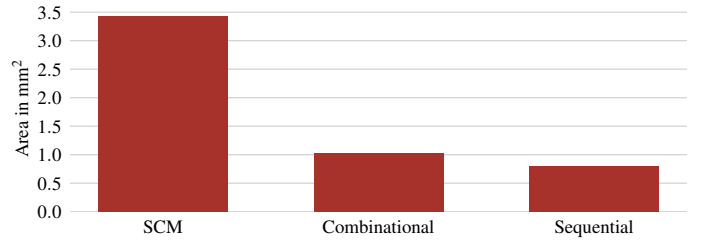


Fig. 7. Breakdown of the area usage of the SCM implementation of the accelerator core in 22 nm technology. The majority of the area is used by the standard cell memories, which are used to store feature maps and weight kernels. Clock area is negligibly small, due to deliberate low clock speeds and hierarchical clock gating

TABLE III
LAYER ARCHITECTURE OF THE TESTED CNN

Layer	Input Dim	Op	Kernel	Padding
2D Convolution	$126 \times 32 \times 32$	297 MOp	3×3	(1,1)
2D Convolution	$128 \times 32 \times 32$	302 MOp	3×3	(1,1)
2D Convolution	$128 \times 32 \times 32$	302 MOp	3×3	(1,1)
Max Pooling	$128 \times 32 \times 32$	-	2×2	(0,0)
2D Convolution	$128 \times 16 \times 16$	75.5 MOp	3×3	(1,1)
2D Convolution	$128 \times 16 \times 16$	75.5 MOp	3×3	(1,1)
Max Pooling	$128 \times 16 \times 16$	-	2×2	(0,0)
2D Convolution	$128 \times 8 \times 8$	18.9 MOp	3×3	(1,1)
2D Convolution	$128 \times 8 \times 8$	18.9 MOp	3×3	(1,1)
Max Pooling	$128 \times 8 \times 8$	-	2×2	(0,0)
2D Convolution	$128 \times 4 \times 4$	4.7 MOp	3×3	(1,1)
Avg Pooling	$128 \times 4 \times 4$	-	4×4	(0,0)
Fully connected	128	2.6 KOp	-	-
Total	-	1.1 GOp	-	-

thermometer encoding and the ternary thermometer encoding for input encoding. The network architecture is shown in Table III.

Each convolutional layer is followed by a batch normalization layer and a Hardtanh activation [70] layer. For the quantized versions of the network, the activation layer is followed by a ternarization layer. The preceding convolutional layer, batch normalization layer and Hardtanh activation layer are merged into a single Fused Convolution layer. Any succeeding pooling layers are then merged as well. The reason for using Hardtanh activations over, for example, the more popular ReLU activation which is also usually used in BNNs is the inclusion of all three ternary values in the range of the function. We further found that the Hardtanh activation converged much more reliably than the ReLU activation for the experiments we ran. We have tested networks with depthwise-separable convolutions in place of standard convolutions but have found that accuracy decreases substantially when ternarizing these networks, which is in line with the results in [37] and [71]. Further, depthwise-separable convolutions require twice the feature map data movement, while performing fewer operations overall. Since CUTIE's architecture greatly reduces the cost of the elementary multiply and add operations, the cost of accessing local buffers is relatively high. Hence, layers that have been optimized in a traditional setting to minimize the number of operations are not guaranteed to be energy efficient.

The approach for training the networks taken in this work is based on the INQ algorithm [32]. Training is done in full-precision for a certain number of epochs, after which a

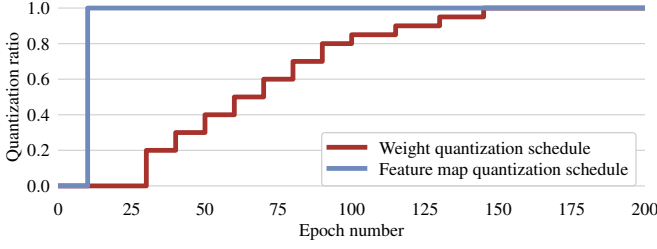


Fig. 8. Quantization schedule for the presented network. Weights and feature map pixels are quantized separately, using different schedules. The weight quantization schedule uses a decaying step size, which starts at 20%, decreases to 10% and finishes with 5% of all weights.

pre-defined ratio of all weights are quantized according to a quantization schedule. These two steps are iterated until all weights are quantized. One degree of freedom in this algorithm is the order in which the weights are quantized, called the quantization strategy. We evaluated three quantization strategies for their impact on accuracy, and sparsity, which is linked to energy efficiency for execution on the proposed architecture. The strategies evaluated in this work are the following:

- **Magnitude:** Weights are sorted in descending order by their absolute value
- **Magnitude-Inverse:** Weights are sorted in ascending order by their absolute value
- **Zig-Zag:** Weights are sorted by taking the remaining smallest and largest values one after another.

For both the ternarized and binarized versions, the weights were quantized using the quantization schedule shown in Figure 8. The CIFAR-10 dataset was used for training and the CIFAR-10 test data set was used for all evaluations. The network was trained using the ADAM optimizer [72] over a total of 200 epochs.

B. Evaluation Setup

In addition to the quantized network, a testbench was implemented to simulate the cycle-accurate behavior of the accelerator core. The testbench generates all necessary signals to load all weights and feature maps into the accelerator core and load the layer instructions into the layer FIFO. The 22 nm implementations were simulated using annotated switching activities from their respective post-layout netlist to simulate the average power consumption of the accelerator core, including memories, during the execution of each layer. Analogously, the 7 nm implementation was simulated using its post-synthesis netlist. For power simulation purposes, each layer was run separately from the rest of the network. This guarantees that each loading phase is associated with its layer, which is required to properly estimate the energy consumption of a layer. For throughput and efficiency calculations, the following formula for the number of operations in convolutional layers is used:

$$\Gamma = 2 \cdot I_W \cdot I_H \cdot K \cdot K \cdot N_I \cdot N_O$$

where K corresponds to the side length of the convolutional kernel, I_W and I_H are the output features maps' width and

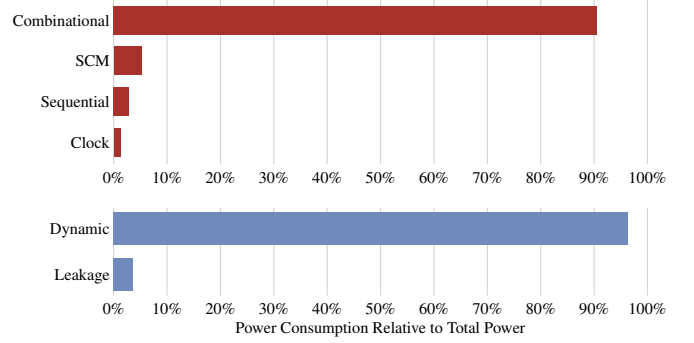


Fig. 9. Power breakdown of the accelerator core implementation in 22 nm technology with SCM-based feature map and weight memories, running the Magnitude-Inverse trained ternary network. The overall power is clearly dominated by combinational cells, where over 90% of the total power is spent.

height, and N_I & N_O are the input and output channel number, respectively. Γ corresponds to the number of additions and multiplications required to compute each output pixel, i.e. operations for pooling and activations are not considered. Furthermore, the runtime of each layer is measured between the loading of the layer instruction and the write operation for the last output feature map pixel.

C. Experimental Results

The energy per operation for the 22 nm implementation using different quantization strategies is shown in Figure 11. The energy efficiency scales almost linearly with the sparsity of the executed network. This trend can be explained by zeros in the adder trees leading to nodes not toggling, which results in lower overall activity.

A breakdown of power consumption by cell type, as well as by dynamic and leakage power is shown in Figure 9. The static power consumption makes up 4.6% of the overall power consumption in the 22 nm implementation, most of which stems from the SCMs. Notably, the power consumption is dominated by combinational cells which underlines the effectiveness of the architecture, since this implies most energy is spent in computations, rather than memory accesses or transfers.

The analysis of the per-layer energy efficiency for both binary and ternary neural networks reveals a sharp peak in the first layer, which can be explained with the structural properties of the thermometer encoding, i.e. the first feature map contains 66.3% zeros on average. Furthermore, with the decreasing number of operations in deeper layers, the energy cost of loading the weights increase in proportion to the energy cost of computations, which explains the decreasing energy efficiency in deeper layers.

The binary thermometer encoding and ternary thermometer encoding were compared for their use with the ternarized network version. The results show that the ternary thermometer encoding provides a small increase between 0.5% and 1.5% in test accuracy, while energy efficiency is kept within 2% of the binary thermometer. Further, the drop in accuracy between the 32-bit full-precision version and the ternary version can be reduced to as little as 3%.

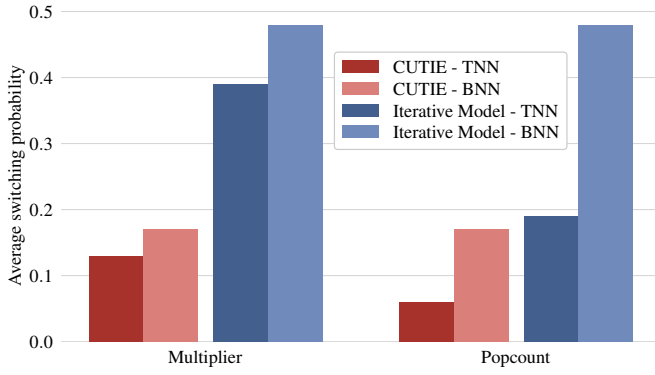


Fig. 10. Overview of the switching probabilities at the multiplier and adder tree input nodes respectively, smaller is better. For the binary case, toggling in the multipliers directly translates to switching activity in the adder trees, while for the ternary case the sparsity of the network reduces switching activity at the adder tree input nodes by $\approx 2\times$. Moreover, the smoothness of feature maps is exploited by unrolling the compute units, which is reflected in a $\approx 3\times$ smaller switching probability compared to an iteratively decomposed model. Best viewed in color.

TABLE IV

IMPACT OF QUANTIZATION STRATEGY ON TEST ACCURACY AND SPARSITY FOR BINARIZED & TERNARIZED NETWORKS ON THE CIFAR-10 DATASET EVALUATED IN THE 22nm SCM IMPLEMENTATION

	Accuracy	Weighty Sparsity	Avg. TOP/s/W
Full-Precision	91%	-	-
Ternary, TT*			
Magnitude	86.5%	7.4%	260 TOP/s/W
Magnitude-Inverse	87.4%	60.7%	392 TOP/s/W
Zig-Zag	88.1%	49.1%	345 TOP/s/W
Ternary, BT*			
Magnitude	85.9%	6.9%	262 TOP/s/W
Magnitude-Inverse	86.8%	60.8%	399 TOP/s/W
Zig-Zag	86.6%	49.2%	342 TOP/s/W
Binary			
Magnitude	83.3%	0%	240 TOP/s/W
Magnitude-Inverse	80.1%	0%	248 TOP/s/W
Zig-Zag	82.8%	0%	229 TOP/s/W

* BT: Binary Thermometer

* TT: Ternary Thermometer

Finally, the ternary network trained with the Magnitude-Inverse quantization strategy using the ternary thermometer encoding was evaluated on the post-synthesis netlist of the 7nm implementation, achieving a peak energy efficiency of 3'140 TOP/s/W in the first layer and an average efficiency of 2'100 TOP/s/W.

D. Comparison of Quantization Strategies

An overview of test accuracy and sparsity for all tested strategies is given for the binarized and ternarized versions in Table IV.

The energy per inference for the most efficient ternary version in 22nm adds up to 2.8μJ, the energy per inference for the best binary version to about 4.4μJ. These results allow three observations: first, the quantization strategy not only impacts the accuracy of the resulting network but also the distribution of weights - the number of zeros for the Magnitude-Inverse strategy is more than 8x higher than for Magnitude, at comparable accuracy. The second observation

is that energy efficiency increases significantly for very sparse networks. The Magnitude-Inverse strategy trains a network that runs 36% more efficiently than the one trained with Magnitude for the ternary case. Lastly, the results imply that the optimal quantization strategy might be different for the binary and ternary case. Most importantly, for all training experiments we have run, we have found that ternary neural networks consistently outperform their binary counterparts on the CUTIE architecture by a considerable margin, both in terms of accuracy, with 5% higher test accuracy, as well as in terms of energy efficiency, with 36% lower energy per inference.

E. Exploiting Feature Map Smoothness

By fully unrolling the compute units with respect to the feature map channels and weights, we reduce switching activity in the adder tree of the compute units by an average of 66.6% with respect to architectures that use an output-stationary approach and iterative decomposition. Iteratively decomposed architectures require the accelerator to compute partial results on partial feature maps and weight kernels. The typical approach to implement this is tiling the feature map and weight kernels in the input channel direction, and switch the weight and feature map tiles every cycle. This leads to much higher switching activity.

In the ternary case, an input node of the adder tree switches when the corresponding weight value is non-zero and the feature map value changes. Calculating the mean number of value switches between neighboring pixels, we found that the binary feature map pixels have an average Hamming distance of 44 out of 256 bit and the ternary feature map pixels have an average pixel-to-pixel Hamming distance of 33 out of 256 bit following the 3-ary encoding of CUTIE. It exploits this fact by keeping the weights fixed for the execution of a full layer, which eliminates switching activity due to changing the weight tile while a previous feature map tile is scheduled. To quantify this effect, we analyzed the switching activity of the presented network trained with all quantization strategies on an output-stationary iterative architecture model, taking into account the network weights as well. Figure 10 shows the occurring switching activity for CUTIE versus a model with $2\times$ iterative decomposition for the binary Magnitude and ternary Magnitude-Inverse trained networks.

F. Comparison of Binary and Ternary Neural Networks

Since the set of ternary values includes the set of binary values, a superficial comparison between binary and ternary neural networks on the proposed accelerator architecture is fairly straight-forward, as binary neural networks can be run on the accelerator as-is. To fairly compare, however, it is important to discount certain contributions that only appear because the accelerator core supports ternary operations. Most importantly, the overhead in memory storage, accesses, encoding, and decoding should be subtracted, as well as the energy spent in the second popcount module. To apply these considerations on the architecture, the following simplifications are made:

- The power used for memory accesses is divided by 1.6.

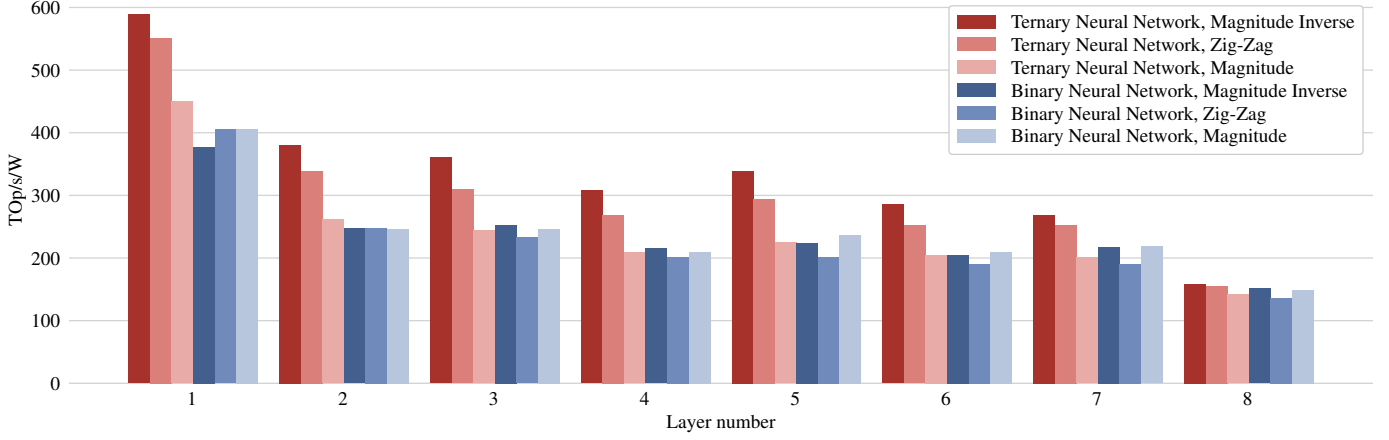


Fig. 11. Energy efficiency simulation results on the CIFAR-10 test dataset for the binarized & ternarized networks comparing the different quantization strategies using the GF 22 nm post-layout power simulation data. Notably, the energy efficiency per operation increases with increasing sparsity of the weight kernels as shown in table IV.

- The power used in the popcounts of the compute units is halved.
- The power used for encoding and decoding is subtracted.

While these reductions do not account for all differences between the ternary and a binary implementation of the accelerator, they give a reasonably close estimate, considering that the power spent in popcounts, memories and encoding & decoding modules accounts for around 80% of the total power budget. Adding up the reductions, an average of around 30% should be subtracted from the measured values of the GF 22 nm SCM implementation to get an estimate for the energy efficiency of a purely binary version of the accelerator. Even including this discount factor into all calculations, the energy of the binary neural network would be reduced to around 3 μ J, which is slightly higher than the ternary version. Taking into account that the achieved accuracy for the ternary neural network comes in at around 88% while the binary version achieves around 83%, the ternary implementation is both more energy-efficient and more accurate in terms of test accuracy than the binary version.

G. Comparison with the State-of-the-Art

A comparison of our design with similar accelerators cores is shown in Table V. The implementation in TSMC 7 nm technology outperforms even the most efficient digital binary accelerator design, implemented in comparable Intel 10 nm technology as reported by Knag et al. [27], by a factor of at least $3.4\times$ in terms of energy efficiency per operation and $5.9\times$ in terms of energy per inference as well as the most efficient mixed-signal design as reported by Bankman et al. [25], requiring a factor of $4.8\times$ less energy per inference.

For a fairer comparison to other state-of-the-art accelerators, we also report post-layout simulation results in GF 22 nm technology, which similarly outperforms comparable implementations as reported in Moons et al. [21] by a factor $2.5\times$, both in terms of peak efficiency as well as average efficiency per operation. The more practical comparison between the energy per inference on the same data set reveals that our design outperforms all other designs by an even larger margin, i.e. by at least $4.8\times$, while even increasing the inference accuracy with respect to all other designs. However, our design is less efficient in terms of throughput per area compared to other state-of-the-art designs. This is a deliberate design

TABLE V
COMPARISON OF THE PROPOSED ARCHITECTURE TO STATE-OF-THE-ART ACCELERATORS

	[19]	[21]	[25]	[27]	[23]	This work		
Computation Method	digital	digital	mixed	digital	analog	digital	digital	digital
Weight Precision	binary	binary	binary	binary	ternary	ternary	ternary	ternary
Activation Precision	binary	binary	binary	binary	ternary	ternary	ternary	ternary
Memory Implementation	SCM	SRAM	SRAM	SCM	SRAM	SRAM	SCM	SCM
Technology	22 nm	28 nm	28 nm	10 nm	32 nm	22 nm	22 nm	7 nm
Core Area [mm^2]	0.7	1.4	5.76	0.39	1.96	7.5	7.5	1.2 ^b
Core Voltage [V]	0.4	0.66	0.6	0.37	-	0.65	0.65	0.65
Peak Throughput [TOP/s]	0.3	2.8	-	160	114	16	16	16
Peak Core Energy Efficiency [TOP/s/W]	223	230	-	617	-	457	589	3'140
Average Core Energy Efficiency [TOP/s/W]	36	145	772	617	127	305	392	2'100
Accuracy on CIFAR-10	87%	86%	85.6%	86% ^a	-	88%	88%	88%
Energy per Inference on CIFAR-10 [μ J] (excl. I/O)	1.3–7.3	13.86	2.61	3.2	-	3.6	2.8	0.52

^a: uses same network as [21] ^b: expected value at 0.75 cell layout density

choice, which is due to the unrolled architecture of CUTIE.

VI. CONCLUSION

In this work, we have presented three key ideas to increase the core efficiency of ultra-low bit-width neural network accelerators and evaluated their impact in terms of energy per operation by combining them in an accelerator architecture called CUTIE. The key ideas are: 1) completely unrolling the data path with respect to all feature map and filter dimensions to reduce data transfer cost and switching activity by making use of spatial feature map smoothness, 2) moving the focus from binary neural networks to ternary neural networks to capitalize on the inherent sparsity and 3) tuning training methods to increase sparsity in neural networks at iso-accuracy. Their combined effect boosts the core efficiency of digital binary and ternary accelerator architectures and contribute to what is to the best of our knowledge the first digital accelerator to surpass POPs/W energy efficiency for neural network inference.

Future work will focus on extending the core architecture to enable efficient computation of different layers and integrating the accelerator core into a sensor system-on-chip.

ACKNOWLEDGEMENT

The authors would like to thank *armasuisse Science & Technology* for funding this research. This project was supported in part by the EU's H2020 Programme under grant no. 732631 (OPRECOMP).

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, p. 1097–1105.
- [2] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey and benchmarking of machine learning accelerators," in *Proc. IEEE High Performance Extreme Computing Conference*, 2019.
- [3] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.
- [4] M. Gautschi, P. D. Schiavone, A. Traber, I. Loi, A. Pullini, D. Rossi, E. Flamand, F. K. Gurkaynak, and L. Benini, "Near-Threshold RISC-V Core With DSP Extensions for Scalable IoT Endpoint Devices," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 25, no. 10, pp. 2700–2713, 2017.
- [5] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," in *Proc. International Conference on Learning Representations*, 2016.
- [6] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Proc. NIPS*, 2015, p. 1135–1143.
- [7] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," *arXiv:1602.07360*, 2016.
- [8] L. Cavigelli, G. Rutishauser, and L. Benini, "Ebpc: Extended bit-plane compression for deep neural network inference and training accelerators," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 4, pp. 723–734, 2019.
- [9] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *Neural Information Processing Systems*, 2015, pp. 3123–3131.
- [10] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European Conference on Computer Vision*, 2016.
- [11] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 4107–4115.
- [12] F. Li, B. Zhang, and B. Liu, "Ternary weight networks," *NIPS Workshop on Efficient Methods for Deep Neural Networks*, 2016.
- [13] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," in *Proc. ICLR*, 2017.
- [14] X. Lin, C. Zhao, and W. Pan, "Towards accurate binary convolutional neural network," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 345–353.
- [15] H. Alemdar, V. Leroy, A. Prost-Boucle, and F. Pétrot, "Ternary neural networks for resource-efficient ai applications," in *2017 International Joint Conference on Neural Networks*, 2017, pp. 2547–2554.
- [16] H. Qin, R. Gong, X. Liu, X. Bai, J. Song, and N. Sebe, "Binary neural networks: A survey," *Pattern Recognition*, p. 107281, Feb 2020.
- [17] A. Mishra, E. Nurvitadhi, J. J. Cook, and D. Marr, "WRPN: Wide reduced-precision networks," in *Proc. ICLR*, 2018.
- [18] J. Choi, P. I.-J. Chuang, Z. Wang, S. Venkataramani, V. Srinivasan, and K. Gopalakrishnan, "Bridging the accuracy gap for 2-bit quantized neural networks (qnn)," 2018.
- [19] R. Andri, G. Karunaratne, L. Cavigelli, and L. Benini, "Chewbaccann: A flexible 223 tops/w bnn accelerator," 2020.
- [20] R. Andri, L. Cavigelli, D. Rossi, and L. Benini, "Yodann: An ultra-low power convolutional neural network accelerator based on binary weights," in *Proc. IEEE Computer Society Annual Symposium on VLSI*, 2016, pp. 236–241.
- [21] B. Moons, D. Bankman, L. Yang, B. Murmann, and M. Verhelst, "Binareye: An always-on energy-accuracy-scalable binary cnn processor with all memory on chip in 28nm cmos," in *Proc. IEEE CICC*, 2018.
- [22] Y. Chen, T. Yang, J. Emer, and V. Sze, "Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 292–308, 2019.
- [23] S. Jain, S. Gupta, and A. Raghunathan, "Tim-dnn: Ternary in-memory accelerator for deep neural networks," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 28, pp. 1567–1577, 2020.
- [24] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, "A 64-tile 2.4-mb in-memory-computing cnn accelerator employing charge-domain compute," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, 2019.
- [25] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8 μ j/86% cifar-10 mixed-signal binary cnn processor with all memory on chip in 28-nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 158–172, 2019.
- [26] M. Klachko, M. R. Mahmoodi, and D. Strukov, "Improving noise tolerance of mixed-signal neural networks," in *2019 International Joint Conference on Neural Networks*, 2019, pp. 1–8.
- [27] P. C. Knag, G. K. Chen, H. E. Sumbul, R. Kumar, M. A. Anders, H. Kaul, S. K. Hsu, A. Agarwal, M. Kar, S. Kim, and R. K. Krishnamurthy, "A 617 tops/w all digital binary neural network accelerator in 10nm finfet cmos," in *2020 IEEE Symposium on VLSI Circuits*, 2020, pp. 1–2.
- [28] B. Moons, D. Bankman, and M. Verhelst, *BINAREYE: Digital and Mixed-Signal Always-On Binary Neural Network Processing*. Cham: Springer International Publishing, 2019, pp. 153–194.
- [29] Q. Hu, P. Wang, and J. Cheng, "From hashing to cnns: Training binaryweight networks via hashing," in *AAAI Conference on Artificial Intelligence*, 2018.
- [30] G. Cerutti, R. Andri, L. Cavigelli, E. Farella, M. Magno, and L. Benini, "Sound event detection with binary neural networks on tightly power-constrained iot devices," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, 2020, pp. 19–24.
- [31] L. Deng, P. Jiao, J. Pei, Z. Wu, and G. Li, "Gxnor-net: Training deep neural networks with ternary weights and activations without full-precision memory under a unified discretization framework," *Neural networks : the official journal of the International Neural Network Society*, vol. 100, pp. 49–58, 2018.
- [32] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, "Incremental network quantization: Towards lossless cnns with low-precision weights," in *Proc. ICLR*, 2017.
- [33] A. Bulat and G. Tzimiropoulos, "Xnor-net++: Improved binary neural networks," in *British Machine Vision Conference*, 2019.
- [34] M. Spallanzani, L. Cavigelli, G. P. Leonardi, M. Bertogna, and L. Benini, "Additive Noise Annealing and Approximation Properties of Quantized Neural Networks," pp. 1–18, 2019. [Online]. Available: <http://arxiv.org/abs/1905.10452>
- [35] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," 2016.

- [36] B. Zhuang, C. Shen, M. Tan, L. Liu, and I. Reid, "Structured binary neural networks for accurate image classification and semantic segmentation," in *Proc. IEEE/CVF CVPR*, 2019, pp. 413–422.
- [37] H. Phan, D. Huynh, Y. He, M. Savvides, and Z. Shen, "Mobinet: A mobile binary network for image classification," in *2020 IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 3442–3451.
- [38] A. Byerly, T. Kalganova, and I. Dear, "A Branching and Merging Convolutional Network with Homogeneous Filter Capsules," 2020. [Online]. Available: <http://arxiv.org/abs/2001.09136>
- [39] L. Deng, P. Jiao, J. Pei, Z. Wu, and G. Li, "GXNOR-Net: Training deep neural networks with ternary weights and activations without full-precision memory under a unified discretization framework," *Neural Networks*, vol. 100, pp. 49–58, 2018.
- [40] X. Sun, S. Yin, X. Peng, R. Liu, J. S. Seo, and S. Yu, "XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks," *Proceedings of the 2018 Design, Automation and Test in Europe Conference and Exhibition, DATE 2018*, vol. 2018-Janua, pp. 1423–1428, 2018.
- [41] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big Transfer (BiT): General Visual Representation Learning," 2019. [Online]. Available: <http://arxiv.org/abs/1912.11370>
- [42] P. Yin, S. Zhang, J. Lyu, S. Osher, Y. Qi, and J. Xin, "Binaryrelax: A relaxation approach for training deep neural networks with quantized weights," *SIAM Journal on Imaging Sciences*, vol. 11, no. 4, pp. 2205–2223, 2018.
- [43] S. Darabi, M. Belbahri, M. Courbariaux, and V. P. Nia, "BNN+: Improved Binary Network Training," pp. 1–10, 2018. [Online]. Available: <http://arxiv.org/abs/1812.11800>
- [44] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy: FixEfficientNet," pp. 12–16, 2020. [Online]. Available: <http://arxiv.org/abs/2003.08237>
- [45] L. Cavigelli and L. Benini, "RPR: Random Partition Relaxation for Training: Binary and Ternary Weight Neural Networks," 2020. [Online]. Available: <http://arxiv.org/abs/2001.01091>
- [46] J. Faraone, N. Fraser, G. Gambardella, M. Blott, and P. H. W. Leong, "Compressing low precision deep neural networks using sparsity-induced regularization in ternary networks," 2017.
- [47] A. Marban, D. Becking, S. Wiedemann, and W. Samek, "Learning sparse & ternary neural networks with entropy-constrained ternarization (ec2t)," 2020.
- [48] R. Ding, T.-W. Chin, Z. Liu, and D. Marculescu, "Regularizing activation distribution for training binarized deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [49] V. Sze, Y. Chen, T. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
- [50] L. Cavigelli, M. Magno, and L. Benini, "Accelerating real-time embedded scene labeling with convolutional networks," in *Proc. ACM/IEEE/EDAC Design Automation Conference*, 2015.
- [51] Y. Chen, Y. Xie, L. Song, F. Chen, and T. Tang, "A survey of accelerator architectures for deep neural networks," *Engineering*, vol. 6, no. 3, pp. 264 – 274, 2020.
- [52] J. Fowers, K. Ovtcharov, M. Papamichael, T. Massengill, M. Liu, D. Lo, S. Alkalay, M. Haselman, L. Adams, M. Ghandi, S. Heil, P. Patel, A. Sapek, G. Weisz, L. Woods, S. Lanka, S. K. Reinhardt, A. M. Caulfield, E. S. Chung, and D. Burger, "A configurable cloud-scale dnn processor for real-time ai," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, 2018, pp. 1–14.
- [53] Y. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.
- [54] N. P. J. et al., "In-datacenter performance analysis of a tensor processing unit," in *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture*, 2017, pp. 1–12.
- [55] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "14.5 en-vision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi," in *2017 IEEE International Solid-State Circuits Conference*, 2017, pp. 246–247.
- [56] R. Andri, L. Cavigelli, D. Rossi, and L. Benini, "Hyperdrive: A multi-chip systolically scalable binary-weight cnn inference engine," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 309–322, 2019.
- [57] A. D. Mauro, F. Conti, P. D. Schiavone, D. Rossi, and L. Benini, "Always-on 674μ w@4gop/s error resilient binary neural networks with aggressive sram voltage scaling on a 22-nm iot end-node," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 11, pp. 3905–3918, 2020.
- [58] L. Cavigelli and L. Benini, "Extended bit-plane compression for convolutional neural network accelerators," in *Proc. IEEE AICAS*, 2019, pp. 279–283.
- [59] S. Okumura, M. Yabuuchi, K. Hijioka, and K. Nose, "A ternary based bit scalable, 8.80 tops/w cnn accelerator with many-core processing-in-memory architecture with 896k synapses/mm²," in *2019 Symposium on VLSI Circuits*, 2019, pp. C248–C249.
- [60] K. Ando, K. Ueyoshi, K. Orimo, H. Yonekawa, S. Sato, H. Nakahara, M. Ikebe, T. Asai, S. Takamaeda-Yamazaki, T. Kuroda, and M. Moto-mura, "Brein memory: A 13-layer 4.2 k neuron/0.8 m synapse binary/ternary reconfigurable in-memory deep neural network accelerator in 65 nm cmos," in *2017 Symposium on VLSI Circuits*, 2017, pp. 24–25.
- [61] A. Ardakani, Z. Ji, S. C. Smithson, B. H. Meyer, and W. J. Gross, "Learning recurrent binary/ternary weights," in *Proc. International Conference on Learning Representations*, 2019.
- [62] F. Conti, L. Cavigelli, G. Paulin, I. Susmelj, and L. Benini, "Chipmunk: A systolically scalable 0.9 mm², 3.08gop/s/mw @ 1.2 mw accelerator for near-sensor recurrent neural network inference," in *Proc. IEEE Custom Integrated Circuits Conference*, 2018, pp. 1–4.
- [63] B. Moons and M. Verhelst, "A 0.3–2.6 tops/w precision-scalable processor for real-time large-scale convnets," in *2016 IEEE Symposium on VLSI Circuits*, 2016, pp. 1–2.
- [64] S. Sen and A. Raghunathan, "Approximate computing for long short term memory (lstm) neural networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2266–2276, 2018.
- [65] X. Zhou, Z. Du, S. Zhang, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, and Y. Chen, "Addressing sparsity in deep neural networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, no. 10, pp. 1858–1871, 2019.
- [66] Z. Yuan, Y. Liu, J. Yue, Y. Yang, J. Wang, X. Feng, J. Zhao, X. Li, and H. Yang, "Sticker: An energy-efficient multi-sparsity compatible accelerator for convolutional neural networks in 65-nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 2, pp. 465–477, 2020.
- [67] O. Muller, A. Prost-Boucle, A. Bourge, and F. Pétrot, "Efficient decompression of binary encoded balanced ternary sequences," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 27, no. 8, pp. 1962–1966, 2019.
- [68] J. Buckman, A. Roy, C. Raffel, and I. J. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in *International Conference on Learning Representations*, 2018.
- [69] K. Goetschalckx and M. Verhelst, "Breaking high-resolution cnn bandwidth barriers with enhanced depth-first execution," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 323–331, 2019.
- [70] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, 2011.
- [71] H. Phan, Z. Liu, D. Huynh, M. Savvides, K.-T. Cheng, and Z. Shen, "Binarizing mobilenet via evolution-based searching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [72] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.



Moritz Scherer received the B.Sc. and M.Sc. degree in electrical engineering and information technology from ETH Zürich in 2018 and 2020, respectively, where he is currently pursuing a Ph.D. degree at the Integrated Systems Laboratory. His current research interests include the design of ultra-low power and energy-efficient circuits and accelerators as well as system-level and embedded design for machine learning and edge computing applications. Moritz Scherer received the ETH Medal for his Master's thesis in 2020.



Georg Rutishauser received his B.Sc. and M.Sc. degrees in Electrical Engineering and Information Technology from ETH Zürich in 2015 and 2018, respectively. He is currently pursuing a Ph.D. degree at the Integrated Systems Laboratory at ETH Zürich. His research interests include algorithms and hardware for reduced-precision deep learning, and their application in computer vision and embedded systems.



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. After spending another year as a Postdoc at ETH Zürich, he has joined Huawei's Zurich Research Center in Spring 2020. 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 ICDSC conferences in 2013 and 2017, the best

student paper award at the Security+Defense conference in 2016, the ETH Medal for his Ph.D. thesis in 2019, and the Donald O. Pederson best paper award (IEEE TCAD) in 2019.



Luca Benini is the Chair of Digital Circuits and Systems at ETH Zürich 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.