

DLA: Compiler and FPGA Overlay for Neural Network Inference Acceleration

Mohamed S. Abdelfattah, David Han, Andrew Bitar, Roberto DiCecco, Shane O’Connell, Nitika Shanker, Joseph Chu, Ian Prins, Joshua Fender, Andrew C. Ling, Gordon R. Chiu

Programmable Solutions Group, Intel

Toronto, Canada

{firstname.lastname}@intel.com

Abstract—Overlays have shown significant promise for field-programmable gate-arrays (FPGAs) as they allow for fast development cycles and remove many of the challenges of the traditional FPGA hardware design flow. However, this often comes with a significant performance burden resulting in very little adoption of overlays for practical applications. In this paper, we tailor an overlay to a specific application domain, and we show how we maintain its full programmability without paying for the performance overhead traditionally associated with overlays. Specifically, we introduce an overlay targeted for deep neural network inference with only ~1% overhead to support the control and reprogramming logic using a lightweight very-long instruction word (VLIW) network. Additionally, we implement a sophisticated domain specific graph compiler that compiles deep learning languages such as Caffe or Tensorflow to easily target our overlay. We show how our graph compiler performs architecture-driven software optimizations to significantly boost performance of both convolutional and recurrent neural networks (CNNs/RNNs) – we demonstrate a 3 \times improvement on ResNet-101 and a 12 \times improvement for long short-term memory (LSTM) cells, compared to naïve implementations. Finally, we describe how we can tailor our hardware overlay, and use our graph compiler to achieve ~900 fps on GoogLeNet on an Intel Arria 10 1150 – the fastest ever reported on comparable FPGAs.

I. INTRODUCTION

Creating custom high-performance hardware designs on field-programmable gate arrays (FPGAs) is difficult and time-consuming when compared to software-programmable devices such as CPUs. A hardware designer must describe their system in a cycle-accurate manner, and worry about low-level hardware considerations such as timing closure to memory interfaces. Over the past decade, significant progress has been made in easing the use of FPGAs through high-level languages such as OpenCL, making it easier to implement high-performance designs [9]. However, even when using high-level design, one must still carefully describe an efficient *parallel* hardware architecture that leverages the FPGA’s capabilities such as the massive on-chip memory bandwidth or configurable multiplier blocks. Additionally, the designer must optimize both area and frequency through long compilations to realize performance gains versus other programmable platforms. Compared to writing a software algorithm targeting a CPU, designing for FPGAs is still drastically more difficult. Our goal in this paper is to present a software-programmable hardware overlay on FPGAs to realize the ease-of-use of software programmability and the efficiency of custom hardware design.

We introduce a domain specific approach to overlays that leverages both software and hardware optimizations to achieve state-of-the-art performance on the FPGA for neural network (NN) acceleration. For hardware, we partition configurable parameters into runtime and compile time parameters such that you can tune the architecture for performance at compile time, and program the overlay at runtime to accelerate different NNs. We do this through a lightweight very-long instruction word (VLIW) network that delivers full reprogrammability to our overlay without incurring *any* performance or efficiency overhead (typical overlays have large overhead [4]). Additionally, we create a flexible architecture where only the core functions required by a NN are connected to a parameterizable interconnect (called Xbar). This avoids the need to include all possible functions in our overlay during runtime; rather, we can pick from our library of optimized kernels based on the group of NNs that are going to run on our system. Our approach is unlike previous work that created hardware that can only run a single/specific NN [1], [7], [8].

On the software side, we introduce an architecture-aware graph compiler that efficiently maps a NN to the overlay. This both maximizes the hardware efficiency when running the design and simplifies the usability of the end application, where users are only required to enter domain specific deep learning languages, such as Caffe or Tensorflow, to program the overlay. Our compiler generates VLIW instructions that are loaded into the FPGA and used for reprogramming the overlay in tens of clock cycles thus incurring no performance overhead. Compared to fixed-function accelerators that can only execute one NN per application run, our approach opens the door to allow for multiple NNs be run consecutively in a single application run [12] by simply reprogramming our overlay instead of recompiling or reconfiguring the FPGA.

The rest of this paper is organized as follows. Section II introduces our hardware architecture. We describe how we target specific NNs using our compile-time parameters and Xbar interconnect. Importantly, we describe our lightweight VLIW network in Section II-A, used for programming the overlay. Next, we describe our NN graph compiler in Section III, and detail some of our architecture-driven optimizations that allow the efficient implementation of NNs on architecture variants of different sizes. Sections IV and V detail how our graph compiler and hardware overlay work together for efficient implementation of CNNs and RNNs. We walk through hard-

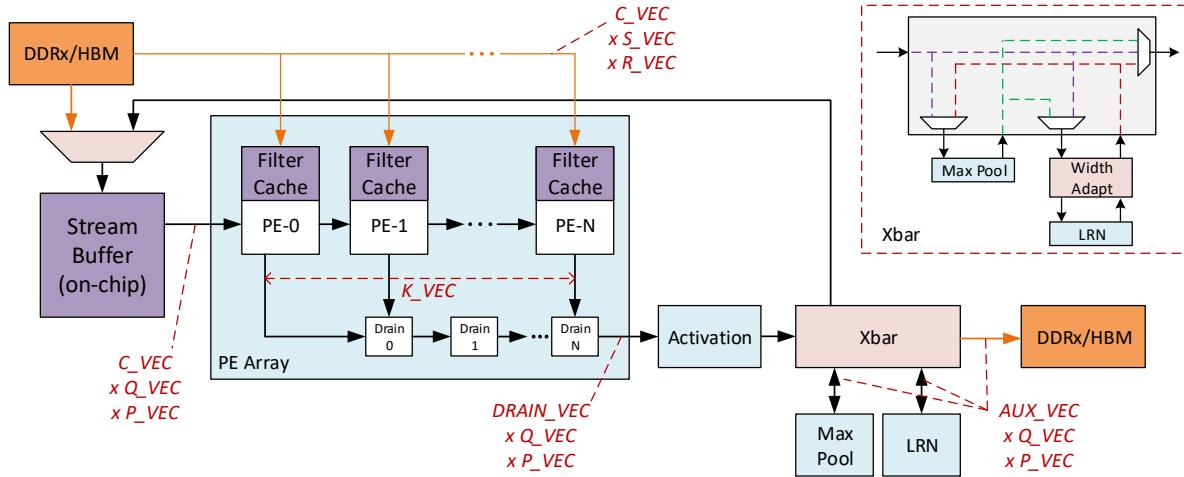


Fig. 1: System-level diagram of our neural network inference accelerator (DLA).

ware and software optimizations in implementing both the ResNet and GoogLeNet CNNs, allowing us to achieve record-setting performance on GoogLeNet. Finally, we discuss the implementation of a long short-term memory (LSTM) cell by simply adding an additional kernel to our overlay, and relying on our graph compiler to mutate the LSTM cell graph to fit within our overlay. In this paper, we refer to our system as “DLA” – our Deep Learning Accelerator.

II. HARDWARE ARCHITECTURE

Our domain specific overlay aims to be general enough to implement any NN, but still remain customizable so that it can be optimized for a specific NN only. Fig. 1 shows an overview of our overlay. At the core of our overlay is a 1D systolic processing element (PE) array that performs dot product operations in each PE to implement general matrix math such as convolutions or multiplications. We omit the discussion of numerics in this paper but we support different floating-point formats such as FP32/16/11/10/9/8 which have been shown to work well with inference [2] – these could be easily modified to support any nascent innovations in data type precisions such as bfloat [15], and other unique fixed or floating point representations, due to the flexible FPGA fabric.

As Fig 1 shows, our Xbar interconnect can augment the functionality of our overlay with different *auxiliary* functions (also referred to as *kernels* in this paper). This section goes through different parts of our hardware architecture and highlights the built-in compile-time flexibility and run-time programmability of our overlay.

A. VLIW Network

To implement a NN on DLA, our graph compiler breaks it into units called “subgraphs” that fit within the overlay’s buffers and compute elements. For example, with convolutional neural networks (CNNs), a subgraph is typically a single convolution with an optional pooling layer afterwards. We deliver new VLIW instructions for each subgraph to program DLA correctly for the subgraph execution.

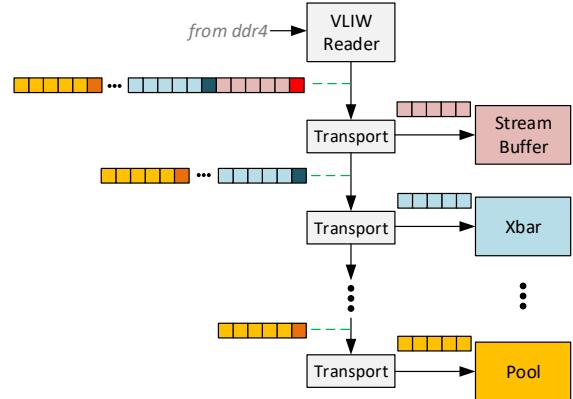


Fig. 2: VLIW network distributes instructions to each kernel.

Our novel VLIW network distributes instructions to each kernel as shown in Fig. 2. The VLIW reader continuously fetches the instructions for the next subgraph from external memory and sends it down an 8-bit unidirectional ring network that is connected to all of the kernels in DLA. The VLIW instruction sequence is divided into different portions for each kernel. A special header packet identifies the kernel, then it is followed by a series of programming instructions that are destined for that kernel. The “Transport” kernels parse the header packet and redirect the instructions that follow to the correct kernel as shown in Fig. 2. The transport kernels also assemble the 8-bit packets into 32-wide instructions for direct kernel consumption.

Our *instructions* are actually counter end values and control flags that are directly loaded into registers within each kernel to govern its operation – this avoids the need for any instruction decode units. For example, the pool kernel receives approximately a dozen instructions: the image height/width/depth, the pool window size, and the type of pooling (maxpool or average pool). Before executing each subgraph, the pool kernel would read each of its 12 instructions serially, consuming 12 clock cycles – this has no material impact on performance that typically takes thousands of cycles. However, it ensures that the entire VLIW network can remain

only 8 bits wide, with a minimal area overhead of only ~3000 LUTs – about 1% of an Arria-10 1150 FPGA device as shown in Table I. Adding new auxiliary programmable functions (kernels) to DLA is simple and has little overhead – we extend the VLIW network with an additional transport kernel, and connect that new kernel to the Xbar without affecting existing kernels or instructions.

TABLE I: Area overhead of VLIW network for DLA with 10 kernels at frequency of 450 MHz on Arria 10.

	LUTs	FFs	ALMs
VLIW Reader	1832	1841	1473
Transport	126	139	73
Total	3092	3231	2046

B. Xbar Interconnect

Machine learning is a fast-developing field – we are increasingly seeing new functions implemented by the machine learning research community. For example, new activation functions are constantly being evaluated such as “Swish” [10]. A quick look at Tensorflow shows that there more than 100 different layer types that users can experiment with in building different NNs [15]. We aim to use the Xbar for extensibility of DLA such that users can easily add or remove functions to implement different types of NNs.

Fig. 1 shows an example Xbar interconnect used to connect pool/LRN kernels for CNNs. As the diagram shows, the Xbar is actually a custom interconnect built around exactly what is needed to connect the auxiliary kernels. For example, the SqueezeNet graph has no local response normalization (LRN) layers, so we can remove that kernel completely. From a prototxt architecture description, the Xbar (including width adaptation) is automatically created to connect auxiliary kernels. We use width adapters to control the throughput of each auxiliary kernel – for example, we can decrease the width of infrequent kernels such as LRN to conserve logic resources. The interconnection pattern within the Xbar is also customizable based on the order of the auxiliary operations. For example, the AlexNet graph has both MaxPool and LRN layers, but LRN always comes first; whereas the GoogLeNet graph has some layers in which MaxPool precedes LRN, which is supported by adding more multiplexing logic.

To demonstrate the power of our extensible architecture (and compiler which is presented in Section III), we add a single kernel to the Xbar in Section V which extends our architecture to also implement LSTM cells alongside CNNs – this allows implementing video-based RNNs commonly used for gesture recognition for instance [16].

C. Vectorization

To ensure our overlay can be customized to different neural network models and FPGA devices, we support *vectorization*, or degree of parallelism, across different axes. Figure 1 shows some of the degrees of parallelism available in the accelerator, configurable via vectorization. Q_VEC and P_VEC refer to the parallelism in the width and height dimensions, while C_VEC and K_VEC refer to the input/output depth parallelism respectively. Every clock cycle, we process the product of

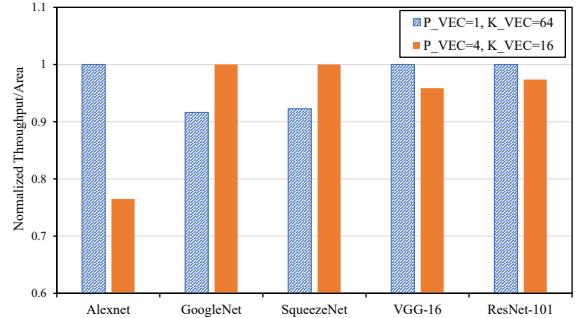


Fig. 3: Throughput/Area on two architectures with different P_VEC and K_VEC vectorization.

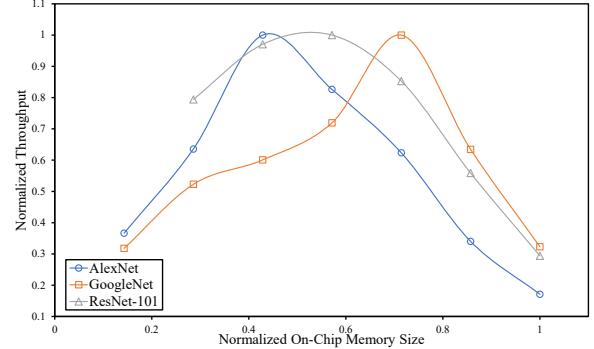


Fig. 4: Impact of stream buffer memory vs. compute tradeoff on AlexNet, GoogleNet and ResNet-101.

{Q_VEC, P_VEC, C_VEC, and K_VEC} feature values in parallel.

Initially, our design was scaled by increasing K_VEC; however, this method of scaling saw diminishing returns, since quantization inefficiencies can become more pronounced as vectorization dimensions increase. For example, if the output depth (K) of a layer is 96, and K_VEC is 64, this will require 2 complete iterations through the PE array, with only 96/128 (75%) useful computations. On the other hand, if K_VEC is 32, the output depth divides perfectly into 3 iterations at 100% efficiency. To mitigate this quantization effect, it is possible to balance the scaling of the design across multiple different dimensions besides just K_VEC (e.g. P_VEC, Q_VEC, C_VEC, etc). The optimal balance of vectorization depends on the graph’s layer dimensions. Figure 3 demonstrates this point by comparing the throughput of two architectures with similar area for different graphs. As the figure shows, the optimal balance of scaling the design between P_VEC and K_VEC varies based on the neural network topology being used. This is an example of how we tune our overlay to get top performance on specific NNs.

D. Stream Buffer and Filter Caches

A single Arria 10 FPGA contains ~4 TB/s on-chip memory bandwidth, interspersed within the FPGA in configurable 20 Kbit memory blocks. This powerful FPGA resource is pivotal in determining the performance of FPGA compute operations – DLA leverages these block RAMs to buffer both activation and filter tensors. As Fig. 1 shows, filters are stored in a double-buffered “filter cache” contained in each PE, allowing the PEs to compute data while filters are

pre-loaded from external memory for the next subgraph. The “stream buffer” is a flexible scratchpad that is used to store intermediate tensors on-chip. Many of our graph compiler passes are dedicated for efficient use of this stream buffer as Section III will show.

When presented with an intermediate tensor larger than the stream buffer or filter caches, our graph compiler *slices* the tensor into multiple pieces that fit within our on-chip caches, and the rest of the pieces are stored in slower off-chip memory, and require higher latency to fetch and compute. To limit this *slicing*, we can increase the size of the stream buffer and/or filter caches, but this decreases the number of RAM blocks available to increase PE array vectorization. Therefore, there is a memory-vs-compute tradeoff for each NN to balance the size of the caches and the number of PEs – Fig. 4 illustrates this tradeoff for different NNs. As the figure shows, a tradeoff that is optimal for one NN can cause 40% or more performance degradation for a second NN.

III. GRAPH COMPILER

The previous section focused on the hardware overlay architecture and how to configure it at compile time to maximize performance for a specific NN graph. This section describes our NN graph compiler that takes advantage of the overlay VLIW instructions to decompose, optimize, and run a NN model on the overlay. The graph compiler breaks down a NN into subgraphs, schedules subgraph execution, and importantly, allocates explicit cache buffers to optimize the use of our stream buffer and filter caches. This section goes through our core compiler “passes” (slicing, scheduling and allocation), and shows examples of how smart graph compilation allows more efficient hardware implementations. Besides these general core passes, our compiler implements more specific algorithms that target and optimize specific NN patterns as we show in the following Sections IV and V.

A. Slicing

To achieve the highest possible throughput for a given DLA architecture it is desirable to size the stream buffer and filter caches in such a way to fit the entire input feature tensor and filter tensor. However, as the resolution of images increases and graph topologies for NNs become deeper, on-chip allocation for these tensors may not be feasible. To overcome this constraint, slices of the input tensor are fetched from external memory into the stream buffer and processed independently by DLA.

The 3D input feature tensor can be sliced along the height, width, or depth to fit in the on-chip stream buffer. When slicing along the width and height, the slices must overlap if the filter window size is greater than 1x1. The graph compiler tries to pick slices that minimize the overlapped computation for the sliced tensor. Alternatively, slicing across the depth does not require overlapped computations, but requires an additive operation to add the results of the depth-wise slices.

To boost performance and minimize the number of DDR4 spillpoints, we enhance our slicing algorithm to slice multiple sequential convolutions together (called “Group Slicing”). Instead of completing all slices within a layer, we compute

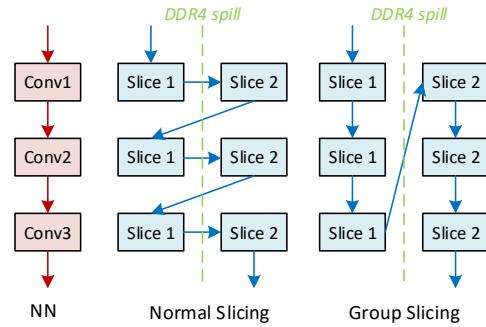


Fig. 5: Group slicing minimizes external memory spillpoints by computing multiple sequential convolutions for each slice.

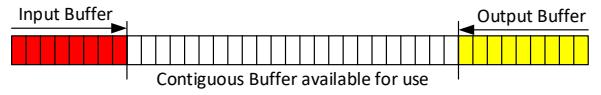


Fig. 6: Double-buffering in the stream buffer.

several sequential convolutions with a single slice using the stream buffer before moving onto the next slice. Fig. 5 illustrates how group slicing reduces the number of external-memory spillpoints for a sample NN. For Resnet101 with image resolution of 1080p (HD), our Group Slicing algorithm improves throughput by 19% compared to simple slicing.

B. Allocation

The allocation pass manages reading and writing from the stream buffer. Allocation calculates the read and write addresses for each slice, and computes the total stream buffer memory used by a graph. One of the main goals is to reduce fragmentation – gaps between allocated memory blocks in the stream buffer. In its most simple operation, the stream buffer is used as a double buffer to store both the input and output of a subgraph. To achieve this double-buffering while reducing fragmentation, the input buffer starts at address 0 and counts up, while the output buffer starts at the end of the stream buffer and counts down. As Fig. 6 shows, this leaves a contiguous space in the middle of the stream buffer that can be used to allocate more data slices in the stream buffer; this is especially useful for graphs that have multiple branches as demonstrated by the GoogLeNet example in Section III-C. Note that the allocation pass must keep track of the *lifetime* of each buffer to be able to free/overwrite its memory in the stream buffer once it is no longer used. Additionally, our allocation pass also assigns addresses in external memory when the stream buffer isn’t large enough, but external memory size is not a problem so it is simply done left-to-right, in the first available space.

C. Scheduling

The DLA compiler partitions NNs into subgraphs where a subgraph is a list of functions that can be chained together and implemented on DLA without writing to a buffer, except at the very end of the subgraph execution – scheduling decides when each subgraph is executed. In the case of early CNN models such as AlexNet [5] or VGG16 [17] there is very little need for a scheduler as there are no decisions to be made on which subgraph to execute next. When considering CNNs

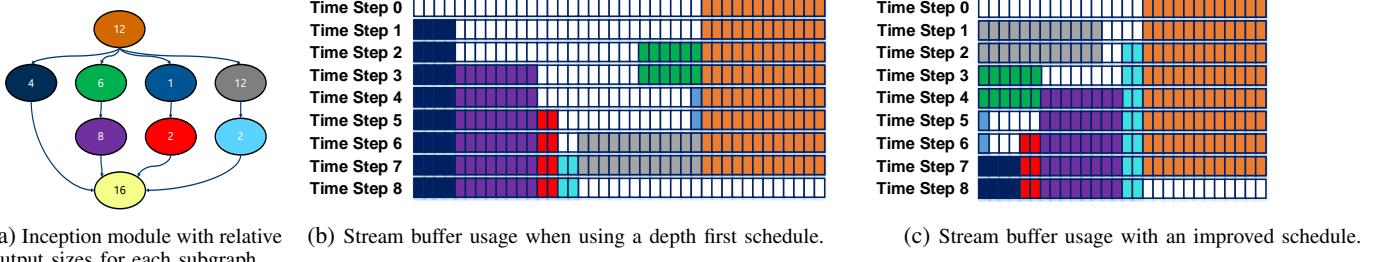


Fig. 7: Scheduling one of the GoogLeNet [6] inception modules.

with branching nodes such as GoogLeNet [6], ResNet [14], or graphs that require slicing, the order of subgraph execution heavily influences the stream buffer size that is required for a given graph to avoid external memory spill points.

Fig. 7a illustrates an example of an inception module from Googlenet, partitioned into DLA subgraphs with the relative output sizes of each subgraph. We show the stream buffer allocation corresponding to two possible schedules of the inception module. Both are depth-first schedules, but in Fig. 7b we start with the leftmost branch, while Fig. 7c starts with the rightmost branch. This simple change in schedule results in a 30% reduction in the size of the required stream buffer for this inception module.

When considering large graphs with many branching nodes that either converge to a single output such as GoogLeNet or graphs that diverge to several outputs such as those used for single-shot multibox detection [12], an exhaustive search of all possible schedules may be infeasible without incurring large compile time penalties. Our scheduling is conducted using a priority queue based approach, where the cost of executing a given node is determined by the ratio of its output size to its *effective* input size (the size of the input multiplied by the number of users of the input tensor). This approach allows for the stream buffer savings of Fig. 7c to be achieved, with minimal impact on the compiler runtime.

IV. CNN IMPLEMENTATION

This section focuses on 2 popular CNNs: ResNet [14] and GoogLeNet [6]. We explain different hardware/software co-optimizations that are possible because of our runtime reconfigurable and software programmable overlay. This allows us to significantly boost the performance of these CNNs on DLA at runtime with little effort, as we show in our results.

A. ResNet Convolution Merging

ResNet 101 is a large graph that can be targeted for high-definition image resolutions, creating intermediate tensors that require significant slicing to run on the DLA overlay on Arria 10. ResNet is composed of three types of *resmodules*, as shown in Fig. 8. Each type has two convolution branches, merged through an element-wise addition operation (*eltwise*).

We present a resmodule optimization (implemented automatically in our compiler) that eliminates the *eltwise* operation by merging it with the preceding convolution(s). This reduces the total number of arithmetic operations in DLA, and more

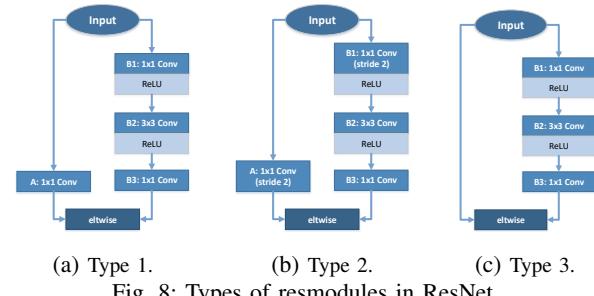
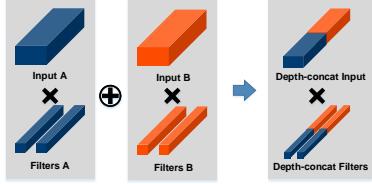


Fig. 8: Types of resmodules in ResNet.

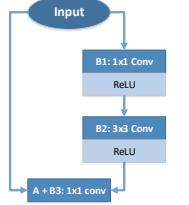
importantly, decreases the number of slices and DDR4 spill-points. Instead of storing intermediate tensors between the convolution and the *eltwise* addition operations, we combine them in a single convolution operation where tensor size is at least half as big as the *eltwise* input.

Consider the computation that produces every output element of the *eltwise* in Type 1 resmodule (Figure 8a) – it is the sum of the corresponding output elements of convolution *A* and *B3*. As illustrated in Figure 9a, this sequence of operations is equivalent to a *single* convolution after input *A* and *B3* (and the corresponding filter *A* and *B3*) are merged depth-wise. This effectively absorbs the *eltwise* addition operation into the dot product operation of the preceding convolutions. Figure 9b shows the Type 1 resmodule after convolution *A* and *B3* are merged with the *eltwise* layer. Since this optimization converts the explicit *eltwise* operations into a convolution, output *A* and *B3*, which would usually reside in DDR4 or on-chip memory, become intermediate results of the merged convolution and are stored in on-chip registers. This reduction in memory traffic is especially prominent in resmodules, where output *A* and *B3* are of $4 \times$ the size of input *A* and *B3*.

In order for Type-2 and Type-3 resmodules to benefit from this optimization, we convert them to Type 1. For Type 2 (Figure 10a), we push the stride-2 convolution *A* and *B1* upstream to the layer *before* the input. Not only does this convert the resmodule to Type 1, it also cuts the amount of computation in the upstream layer and reduces the input traffic to convolution *A* and *B1*. For Type 3 (Figure 10b), we introduce an *identity* convolution – which creates an identical output tensor from an input tensor – in the left branch.

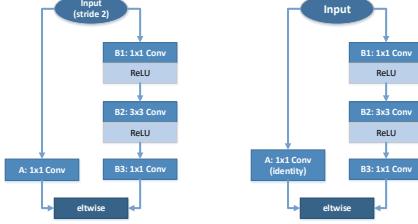


(a) Eltwise elimination.



(b) Optimized Type 1 resmodule.

Fig. 9: Convolution merging optimization.



(a) Type 2.

(b) Type 3.

Fig. 10: Resmodule type conversion to benefit from convolution merging optimization.

B. Non-convolution Primitives

While almost all layers in ResNet are convolutions, there are a couple of exceptions – a single Global Average Pooling (GAP) layer and a single Fully-Connected (FC) layer at the end. This is also true for GoogLeNet where there is a single FC layer, and a single average pooling layer. Given the extremely low frequency of these non-convolution layers (e.g., 2 out of 147 for ResNet 101), it is best to map them to convolutions. In this way, we can reuse the powerful convolution engine (PE array) instead of adding dedicated auxiliary kernels that would be under-utilized (over time).

An FC layer performs a multiplication between a vector (input) and a matrix (weights). It can be mapped to a convolution as follows: 1) the 1D FC input of length N is mapped to a 3D convolution input of shape $1 \times 1 \times N$, and 2) the 2D FC weight matrix of shape $N \times M$ is mapped to M 3D convolution filters of shape $1 \times 1 \times N$. With this mapping, the computation of each FC output is assigned to a PE.

Average pooling of window $H \times W$ on a 2D image is equivalent to a 2D convolution with a filter of size $H \times W$. Each filter element is of value $1/(H \times W)$. For a 3D input of depth D , average pooling is applied to *each* 2D input surface, producing the corresponding output surface. In this case, the equivalent convolution filter for the output surface at depth d , is of shape $H \times W \times D$, with all zero filter values except the surface at depth d being the average pooling filter.

C. Sparse Filter Shortening

Even though they save area, the identity and average pooling convolutions introduced in the previous optimizations could come at a high cost to throughput, due to the large but sparse filters involved. For an identity convolution of input

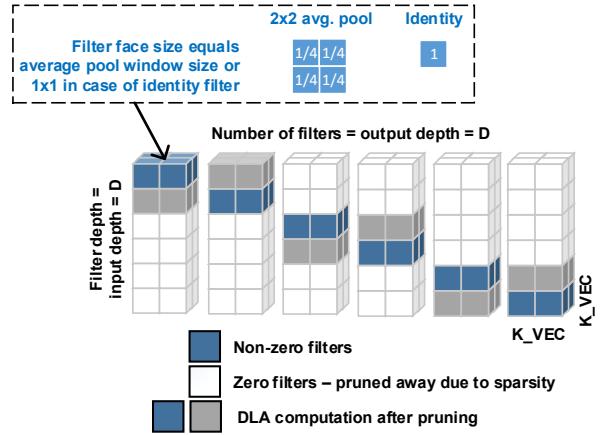


Fig. 11: Sparse filter shortening with identity and average-pooling convolution filters.

and output shape $H \times W \times D$, there are D filters, each of shape $1 \times 1 \times D$. Since each filter is responsible for copying input surface at depth d to the output surface at the same depth, the values of this filter are all zeros except 1 at depth d . Fig. 11 illustrates both the identity and average pooling convolution filters, and how we can leverage their sparsity to conserve operations on DLA. We improve performance by skipping the computation with filter entries that are filled with zeros. Since the PEs process K_VEC filters at a time, we trim the filters size K_VEC to fit perfectly in the PE array. This effectively reduces the filter depth from D to K_VEC , saving both compute time and filter data loading time. We call this optimization *sparse filter shortening*, which can also be applied to the average pooling convolution as shown in Fig. 11, due to the same filter sparsity.

D. 1x1 Filters Optimization

To efficiently compute convolutions using 3x3 filters, the DLA architecture is often tuned to be vectorized in the filter width dimension by setting $S_VEC=3$. Increasing the filter width vectorization increases PE throughput as well as filter prefetch bandwidth for large (eg. 3x3) filters. However, many of the latest CNNs have a mix of 3x3 and 1x1 filters. Convolutions using 1x1 filters do not benefit from filter width vectorization, and thus would achieve low DSP efficiency and filter prefetch bandwidth. To avoid this, the DLA architecture has been optimized for 1x1 filters in two ways. First, the DSPs that would have been used in a 3x3-filter convolution to process the second and third filter values in the filter width direction are instead used to calculate two additional output pixel in a 1x1-filter convolution. This allows the PEs to maintain the same DSP efficiency for both 3x3 and 1x1 filters. Second, the filter prefetch bandwidth added to load a 3-wide filter is used to simply load more 1-wide filters in parallel. Overall, these two optimizations allow DLA to achieve high throughputs through vectorization for 3x3-filter convolutions without suffering any additional quantization loss for 1x1-filter convolutions.

E. Optimization Impact on ResNet

Table II summarizes the impact of each optimization on the throughput of ResNet 101 with 1080p image resolution. The number in each row is the normalized throughput after applying *all* optimizations listed up to this row. Here, we apply the mapping of GAP and FC layers to convolution unconditionally (i.e., in the baseline). The huge speedup of sparse filter shortening comes from the filters of the identity convolutions introduced by convolution merging optimization on Type 3 resmodules which account for 87% of all resmodules in ResNet 101.

TABLE II: Optimization impact on ResNet-101.

Optimization	Relative Throughput
Baseline	1.0
1x1 Filter Opt	1.3
Conv. merging (Type 3)	1.7
Sparse filter shortening	2.8
Group slicing	3.1

F. Optimization Impact on GoogLeNet

Two of the described CNN optimizations are used to improve throughput on GoogleNet: (1) the 1x1 filter optimizations and (2) the average pool mapped to convolution optimization – this allowed DLA to fit a larger PE array instead of wasting dedicated resources on an average-pooling kernel. As shown in Table III, GoogleNet saw a 17% throughput improvement from these two optimizations. The following row in the table shows the throughput improvement from increasing the PE array vectorization (from {P_VEC,K_VEC} = {1,48} to {2,32}). Finally, the last row in the table points to an accurate model of external memory optimizations that will allow DLA to achieve ~900 fps on GoogLeNet on Intel’s Arria 10 1150 device, which to our knowledge, is the most efficient acceleration of GoogLeNet on FPGAs. This optimization entails continuously fetching filters for the next NN layers until the filter cache becomes full instead of limiting filter prefetch only to 1 layer ahead. While this slightly complicates filter prefetch logic, it has a negligible area cost but allows hiding external memory latency when fetching the NN model.

TABLE III: Optimization impact on GoogLeNet.

Optimization	Relative Throughput	Raw Throughput (Intel Arria 10 1150)
Baseline	1.0	469 fps
1x1 Filter Opt	1.1	506 fps
Avg Pool Mapped to Conv	1.2	550 fps
Additional Vectorization	1.7	777 fps
External Memory Opt	1.9	~900 fps

V. LSTM CELL IMPLEMENTATION

LSTM cells are a widely-used variant of RNNs, commonly used in speech recognition [3], translation [18] and motion detection [16]. DLA is designed to be a flexible NN accelerator for all relevant deep learning workloads, including LSTM-based networks. As such, this section discusses how our graph compiler mutates an LSTM cell to map well to the DLA overlay with high performance.

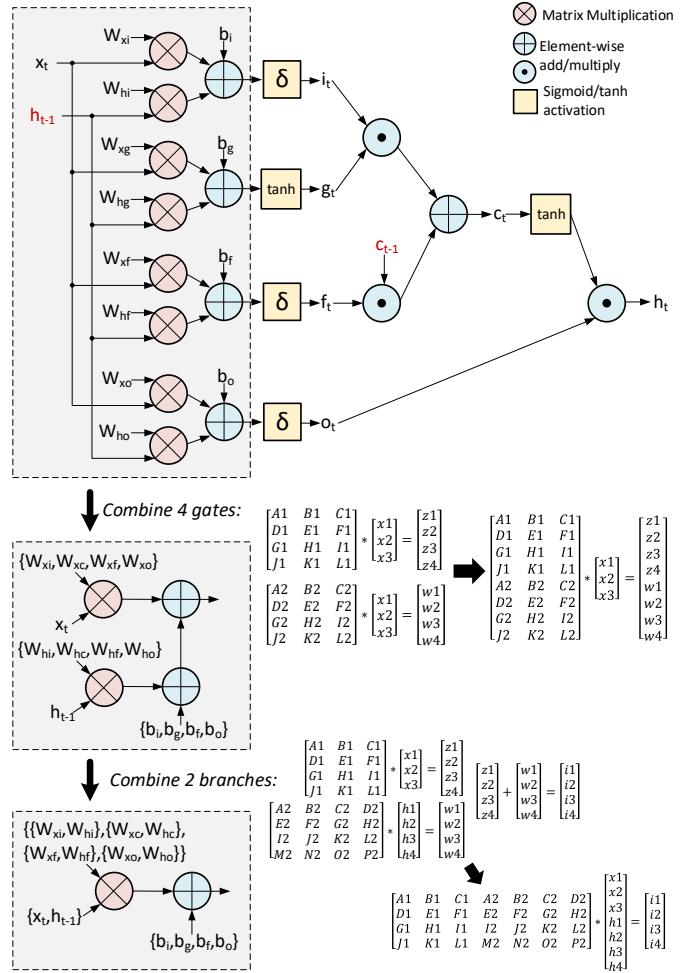


Fig. 12: Graph/Matrix view of an LSTM cell, and how we combine its matrix-multiplications into one big matrix.

A. Mapping an LSTM Cell to DLA

Most of the computation in an LSTM cell occurs in 8 matrix multiplications to compute the 3 LSTM gates (input/forget/output) [11]. Fig. 12 illustrates how we combine those 8 matrices into one big matrix – this reduces DLA execution from 12 subgraphs to a single subgraph which runs at least $\sim 12\times$ faster. First, the 4 matrices that were multiplied by the input/history are each height concatenated as shown in the example in Fig. 12. This is a generic optimization that can be applied to any matrix-vector multiplications that share the same input vector. We end up with two large matrix multiplications, one matrix for the input (x_t), and another for the history (h_{t-1}). Next, we combine those two matrices, and the element-wise addition that follows, into one larger matrix through width concatenation of the matrices, and height concatenation of the input and history vectors as shown in Fig 12. This gives us one large matrix multiplication for the entire LSTM cell. Depending on the LSTM cell size, our compiler may later decide to slice this large matrix if it does not fit on the FPGA as described in Section III-A.

With the combined matrix, each of the LSTM gates are computed one-after-the-other since we compute the matrix

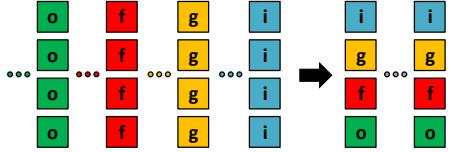


Fig. 13: Matrix row interleaving allows streaming different LSTM gate values simultaneously instead of buffering each gate separately.

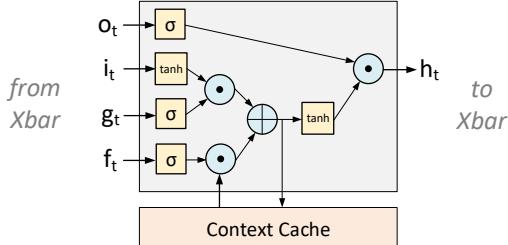


Fig. 14: Streaming LSTM hardware block to compute the element-wise operations of an LSTM cell.

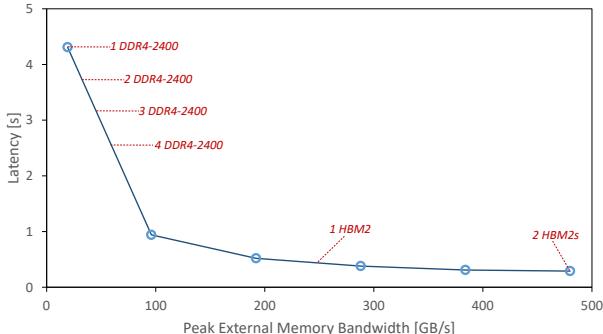


Fig. 15: Latency of an LSTM NN when varying external memory bandwidth.

rows in order. However, this is not FPGA-friendly, as each of the input/forget/output gate values will now need to be buffered (using costly on-chip RAM or slow external memory) so that they can be combined in the second half of the LSTM cell. However, by interleaving the rows of the large matrix (so that the first row contains the filters for the input gate, the second row for the ‘g’ gate, the third row for the forget gate, and the fourth row for the output gate), we can compute one output from each gate in each time step as shown in Fig. 13. This removes the need for buffering large intermediate gate outputs [13], and allows us to directly stream the gate values into the dedicated LSTM hardware block shown in Fig. 14.

This demonstrates the flexibility of the DLA overlay, and the power of our graph compiler in implementing different NNs. By simply attaching the LSTM kernel to the Xbar, we can leverage our powerful multi-precision PE array to compute the matrix-multiplication portion of the LSTM cell, then stream data directly into the dedicated LSTM block.

B. External-Memory-bound RNNs

Non-convolutional neural networks, are effectively a matrix-vector multiplication when computed with batch=1. Most of the applications that use RNNs are real-time applications such as speech/gesture recognition or translation; therefore, they require low-batch and low-latency processing that is ideal

for FPGAs. However, external memory bandwidth is often a bottleneck, since a large matrix has to be fetched from external memory, only to be multiplied with one vector – compute time is lower than memory fetch time so it is *impossible* to hide memory fetch latency. Intel’s Stratix 10 devices have 2 HBM2 devices integrated on some of their boards, providing up to 500 GB/s of peak memory bandwidth – this is 20× higher than a DDR4-2400 memory. In Fig. 15, we look towards the future, and model the performance of a 4-layer stacked LSTM NN (with size of input=output=hidden=2048) used for speech recognition. As the figure shows, with more external memory bandwidth, going from DDR4 to HBM2, the latency for processing a speech segment goes down by more than 5×.

VI. CONCLUSION

We presented a methodology to achieve software ease-of-use with hardware efficiency by implementing a domain specific customizable overlay architecture. We described the hardware tradeoffs involved with NN acceleration, and delved into our graph compiler that maps NNs to our overlay. We then showed that, using both our hardware and software, we can achieve 3× improvement on ResNet 101 HD, 12× on LSTM cells, and 900 fps on GoogLeNet on Intel’s Arria 10 FPGAs. We will further develop DLA to encompass more use-cases such as multi-FPGA deployment [2]. In the future, we also aim to implement similar overlays for different application domains such as genomics, packet processing, compression and encryption to further make FPGAs accessible for high-throughput computation.

REFERENCES

- [1] U. e. a. Aydonat. An opencl™deep learning accelerator on arria 10. In *FPGA*, *FPGA ’17*, pages 55–64, New York, NY, USA, 2017. ACM.
- [2] E. e. a. Chung. Accelerating persistent neural networks at datacenter scale. *HotChips*, 2017.
- [3] A. G. et al. Speech recognition with deep recurrent neural networks. *ICASSP*, 2013.
- [4] A. J. et al. Efficient overlay architecture based on dsp blocks. *FCCM*, 2015.
- [5] A. K. et al. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [6] C. S. et al. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [7] H. Z. et al. A framework for generating high throughput cnn implementations on fpgas. *FPGA*, 2018.
- [8] J. S. et al. Towards a uniform template-based architecture for accelerating 2d and 3d cnns on fpga. *FPGA*, 2018.
- [9] M. A. et al. Gzip on a chip: High performance lossless data compression on fpgas using opencl. *IWOCL*, 2014.
- [10] P. R. et al. Swish: a self-gated activation function. *AISTATS*, 2017.
- [11] S. H. et al. Ese: Efficient speech recognition engine with sparse lstm on fpga. *FPGA*, 2017.
- [12] W. L. et al. Ssd: Single shot multibox detector. *ECCV*, 2016.
- [13] Y. G. et al. Fpga-based accelerator for long short-term memory recurrent neural networks. *ASP-DAC*, 2017.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [15] G. Inc. Tensorflow, 2018.
- [16] K. Murakami and H. Tagushi. Gesture recognition using recurrent neural networks. *SIGCHI*, 1991.
- [17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [18] Y. e. a. Wu. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.