Weight Block Sparsity: Training, Compilers, and Accelerators

Paolo D'Alberto
Taehee Jeong
Akshay Jain
Shreyas Manjunath
Mrinal Sarmah
Samuel Hsu
Nitesh Pipralia

ABSTRACT

We present the main ideas about a vertical system where convolution and matrix multiplication weights can be trained to exploit an 8x8 block sparsity, compilers recognize such a sparsity for both data compaction and computation splitting into threads. If we take a Resnet50, we can reduce the weight by half with little accuracy loss. We can achieve speeds similar to an hypothetical Resnet25. We shall present performance estimates by accurate and complete code generation for a small and efficient set of AIE2 (Xilinx Versal FPGAs).

KEYWORDS

Sparsity, AI, Performance, FPGA, and Tools

ACM Reference Format:
Paolo D'Alberto, Taehee Jeong, Akshay Jain, Shreyas Manjunath, Mrinal Sarmah, Samuel Hsu, and Nitesh Pipralia. 2024. Weight Block Sparsity: Training, Compilers, and Accelerators. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

We shall start by what we mean for block sparsity and for a vertical solution. Block sparsity is an intuitive concept but it is also a little misunderstood. Take a matrix multiplication as in Equation 1

$$\begin{pmatrix} C_0 & C_1 \\ C_2 & C_3 \end{pmatrix} = \begin{pmatrix} A_0 & A_1 \\ A_2 & A_3 \end{pmatrix} \begin{pmatrix} \mathbf{0} & B_1 \\ B_2 & \mathbf{0} \end{pmatrix} \tag{1}$$

This is simply the computation

$$C_0 = A_1B_2;\ C_1 = A_0B_1;\ C_2 = A_3B_2;\ C_3 = A_2B_1 \tag{2}$$

and in general with proper γ_i (i.e., a mask):

$$C_i = \sum_{k=0}^{1} A_{2i+k}\left(\gamma_{2*k+i}B_{2*k+i}\right) \tag{3}$$

Where the matrix $\boldsymbol{B}$ is constant, diagonal, and each submatrix $\boldsymbol{B}_2$ and $\boldsymbol{B}_1$ can split further down and may have even smaller zero blocks. In this work, we chose the basic block of $\boldsymbol{B}_i = 8 \times 8$. It is a great starting point for architectures based on AMD AIE2 products and we support others. For example,

$$\boldsymbol{B} = \dot{\sum_i}\gamma_i\boldsymbol{B}_i, \quad \gamma_i \in \{0,1\} \tag{4}$$

This is a well known data structure in the sparse computation field: We can use Compress Block Row or Column format (CBR). There are standard Matrix Sparse-Matrix Multiplication interfaces and algorithms for CPU and GPUs using this data format (where only one operand is sparse or both) [3, 15].

We explore training techniques (PyTorch and the Keras). The most successful so far is the simplest: we take a pre-trained model, we compute a Γ per layer using a function to determine the blocks more likely to be zeros (Norm) and then we train the model till convergence or accuracy achieved. We take the sparse model and we quantize to 8-bit integer computations by using the Vitis-AI quantizer. The final model is a Xilinx IR quantized model. See Section 3. We have a custom compiler that takes the XIR model and an abstraction of a connected set of AIE2. See Section 4. Given the HW and per layer, the compiler computes the maximum sub-volume computation per core. By heuristics and following a schedule, it computes a memory allocation in mem-tile for input, outputs, and weights. It formats the weights so that to exploit spatial distribution to Mem-tiles and cores into a single compacted tensor. We generate all the explicit communications between DDR, MemTile, and cores. Knowing the subproblem sizes per core and the computation throughput and with a clear specification of what is executed in parallel: we can estimate the execution time per layer and of the entire network with an accuracy closer to a simulation. We shall use this in order to write time estimates for all parts of the computation: we shall show estimates for three CNN models, a eight different AIE designs; see Section **??**.

In the following Section 2, we shall start with a quantitative measure about the advantages of block sparsity in general.

2 BLOCK-SPARSE MATRIX-MATRIX MULTIPLICATION

As mental and practical exercise, consider Γ and Ω two appropriate 0,1 matrices so that for square matrices in $\mathbb{R}^{N\times N}$

$$\boldsymbol{C} = (\Gamma\boldsymbol{A}) * (\Omega\boldsymbol{B})^t \tag{5}$$

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

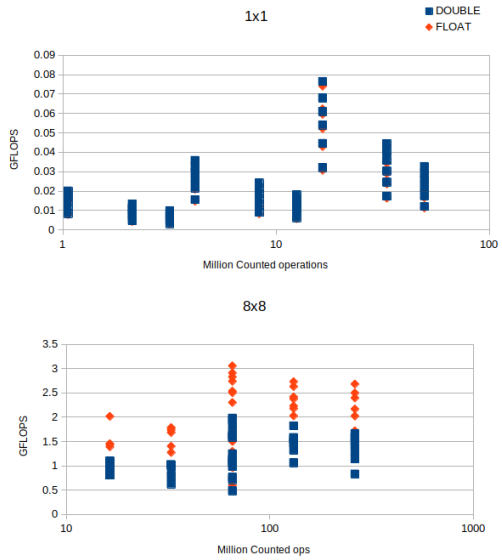

Figure 1: Block 1x1 and 8x8 performance

More precisely, consider non-zero blocks of size $k \times k$ so that

$$C_{i*N+j} = \sum_k (\gamma_{i*N+k} A_{i*N+k})(\dot{\omega}_{j*N+k} \dot{B}_{j*N+k}) \quad (6)$$

Thanks to the sparsity and if we store only non-zeros, then γ_{i*N+k} and $\dot{\omega}_{j*N+k}$ are at the very least contiguous. There will be a meaningful product to compute if and only if $\gamma_{i*N+k} = 1, \dot{\omega}_{j*N+k} = 1$. We merge-sort these vectors. See how the Sparse Sparse Matrix multiplication using Coordinate Block Structure (COO) is [6]. Now, the quantitative part, assume we want to achieve a fixed sparsity (i.e., density) of 50% for a square matrix of size N and we choose the block size $k \times k$. The larger k is, the smaller the overhead will be. The relative performance of the k^3 multiplication is better as k get larger because spatial and temporal locality and optimized code for a constant/parameterized k.

In Figure 1, we present two scatter plots: on the abscissa the effective multiplication-and-addition number, on the ordinate the performance in GFLOPS, when the sparse matrix with dense block is 1x1 and 8x8. Given the same problem, we may use more threads and thus the Jenga of points. We can see that given he same number of effective operations, the block permits better performance and exploits better performance for each precision.

3 BLOCK SPARSITY: TRAINING AND QUANTIZATION

In Convolutional Neural Networks, the two main operations are convolutions/correlations and fully connected layers (matrix multiplication). The block sparsity we are seeking to deploy is not naturally recurring and we must train the network for it.

First, let us clarify block sparsity for convolution weights, then we clarify our training process. A convolution has a weight tensor in four dimension: $\boldsymbol{W} \in \mathbb{R}^{c_{out} \times h \times k \times c_{in}}$. In the hyperplane of the h and k, we can simplify the weight as $\dot{\boldsymbol{W}} \in \mathbb{R}^{c_{out} \times c_{in}}$ and block sparsity can be simply described by a mask $\Gamma\dot{\boldsymbol{W}}$. Although, we speak of a 8×8 of non zeros, this is in practice a $8 \times h \times k \times 8$ block. For the matrix multiply $h = k = 1$, there is no difference from the previous discussions.

3.1 Keras: play

We provide a repository using Keras [5] where we implements the contents of this section: Any one can reproduce and break [7]. We target convolutions only and without quantization. The idea of the framework is simple: we take any model and we create a copy where we enhance the convolution with a (non-trainable) Γ. A convolution will have three parameters (saving the model into a different format). The forward computation is modified so that the weights used for convolution are $\Gamma\boldsymbol{W}$. We assume the backward computation (i.e., gradient) is done automatically from the forward definition. There is no need to change the bias. For example, we take Resnet50 from the keras application repository, we start with a $\Gamma = 1$, and we trained one epoch using imagenet repository [10]. The goal is to choose Γ so that we achieve the required sparsity and to have the minimum loss in accuracy. Using this repository, we can test different ways to train but we skip for lack of space.

3.2 Γ chosen once and full training ahead: PyTorch

Take a convolution with $\Gamma = 1$ and weights $\boldsymbol{W}$. For each γ_i, this will be representative of a block $\boldsymbol{W}_i \in \mathbb{R}^{8 \times h \times w \times 8} \sim \mathbb{R}^{8 \times 8}$. We can choose the $\boldsymbol{W}_i$ using a measure of importance:

- $L_2 = \sqrt{\sum_k w_k^2}$ with w_k elements of $\boldsymbol{W}_i$,
- $L_1 = \sum_k |w_k|$,
- Variance $\sigma^2 = \frac{1}{64} \sum_k (w_k - \mu)^2$ with $\mu = \frac{1}{64} \sum w_k$ or $\frac{1}{N} \sum w_k$ (the whole tensor). In signal processing σ^2 is the power of the signal.

We can then sort them in ascending order and for example choose the first 50% to set to zero. Then we start re-training. We do this for the whole network or for one convolution at a time.

In Table 1, we show the results by using one-time masks and full training: VGG-16, ResNet-50, Inceptionv3 on ImageNet20 (20 classes) and ImageNet1k (1000 classes). See Section 3.2. We use three samples per class for the validation accuracy for ImageNet1k data set; instead, we use 50 samples per class for ImageNet20. Fine-tuning sparse networks on the original ImageNet data-set [10] is expensive. To reduce the training time, we chose 20 classes (from the original 1000 classes) with the least number of images per class in the training data-set and this choice will affect the accuracy because there are fewer sample for re-training.

Classification accuracy on ImageNet1k drops by only 1 - 2% after applying 50% sparsity with a 8×8 block (this is without any quantization). We experiment with different block shapes such as 16×4 and 4×16 on ResNet-50, but the accuracy is slightly worse compared to block of 8×8. Fine-grained sparsity (1×1 block) does not sacrifice any accuracy (i.e., almost any). We use the sparsified models, we quantize them using Vitis AI, and we shall use it for time estimates (i.e., Section ??).

Table 1: Accuracies of the sparsity models

Model	Dataset	Baseline	Sparsity		
		Acc.(%)	block	ratio (%)	Acc.(%)
Inception-v3	ImageNet1k	77.2	8x8	50	75.5
ResNet-50	ImageNet1k	76.7	8x8	50	74.6
VGG-16	ImageNet1k	70.6	8x8	50	69.7
ResNet-50	ImageNet20	96.1	8x8	25	95.1
ResNet-50	ImageNet20	96.1	8x8	50	92.0
ResNet-50	ImageNet20	96.1	8x8	75	87.1
ResNet-50	ImageNet20	96.1	1x1	25	96.0
ResNet-50	ImageNet20	96.1	1x1	50	95.6
ResNet-50	ImageNet20	96.1	1x1	75	93.5
VGG-16	ImageNet20	92.0	8x8	50	89.6
VGG-16	ImageNet20	92.0	1x1	50	92.3
VGG-16	ImageNet20	92.0	1x1	75	91.7

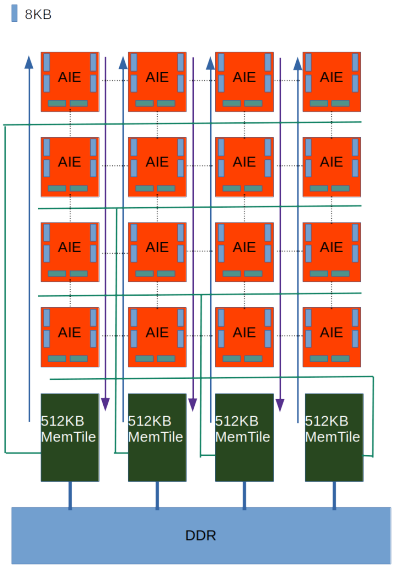

Figure 2: 4x4 AIE representation

4 THE COMPILER AND ITS CODE GENERATION FOR AIE

Within our Xilinx/AMD effort, we can take a PyTorch/Keras model, quantize it using Vitis AI, and create an intermediate representation that we call Xilinx Intermediate Representation (XIR). This is a graph where each node is an operation that reads tensors and writes one tensor. A convolution has one quantized input: we use the tensor layout format BHWC, the tensors are represented in INT8 with a position where the fraction starts (power of two scale). It computes a tensor using the same layout as the input and with a proper scale. The weights and bias are properties of the convolutions, as such they can be tailored; the layout of the weight tensor is $COUT \times h \times w \times CIN$, which is similar to the caffe layout [14] (different from [1]).

For this project, we are at the third generation of a compiler for custom architectures (previously DPUV1 and DPUV3int8 [8, 9]). The main difference is the ability to represent a parameterized block sparsity (block size and overall sparsity ratios) and the capability to split these tensors when we split the computation. All weights are statically prepared into DDR and we move them explicitly towards the inner levels. Inputs and outputs have designated space in DDR. DDR can and it will be used for tensors spills. The memory allocation to Mem-Tile is basically coloring algorithms and some heuristics. In this architecture, we do not allow *streaming* of neither data or weights (because they share space in Mem-Tile). In a previous architecture, we could.

4.1 AIE Hardware abstraction

For this presentation, see Figure 2, we work with a mesh of 4x4 AIE2 cores connected by 4 horizontal and 4 vertical interconnections (we shall presents estimates for square 2x2, .. $i \times i$ .. 8x8 and rectangular shapes are in the works). Each core has 8 banks memories for a total 64KB. About 6 banks are used as input/output/weight buffers and 2 banks are used as temporary space for kernels. Each core can request and send data to its direct neighbors (if aware of connection and control). Double buffering as ping/pong is commonly used for inputs and outputs.

There are four mem-tiles: each 512KB and each can either be connected to one columns and its direct neighbor column, or it can be connected to a row and its neighbor. The total amount of space is 2MB. Memtile is a circular buffer to exploit more flexible allocation, implicitly a 2×2 architecture will have one memtile per column and a total of two (1MB).

A Memtile can broad-cast data per column or per row. We can dedicate a memtile for weights, one for activations, or we can share. We shall show mostly shared configurations. To maximize the computation parallelism, every core will write data per column into memtile: there are different designs that can be used. We shall explore the case where Inputs/Outputs/Weights are evenly distributed across mem-tiles although the distribution is different per tensor.

4.2 Sub Volume, Data compression and Data Movement

The computation is split by column (cores columns): each output tensor is split evenly by width, so each memtile will store different columns by width, each core will compute different output channels, and the computation streams computing the height of the tensor by using core input and output ping/pong. As often as possible, we store the weights in the core and we reuse them as much as possible (unless we need to stream the weight instead). The set of cores is a cohort and symmetric computations are always selected. This means that we do not merge two operations like convolution and max-pool and give three columns to convolution and one column to max-pool.

Before we start describing the memory allocation, we have to explain and compute the size of the sub-volume computation at AIE core. That is, if we have the inputs, output, and weights in memtile, what is the largest computation we can do in the AIE?

The minimum computation is at least one output channel and one row by height. If this is not possible, we try to reduce the size of the width (e.g., shaping the tensor in memtile by using DDR spills) and we can manage to split the input channels (this will require to split the weights accordingly and accumulate). We call W-Split the distribution of tensor by columns in the AIE mesh. We can COUT-split, this will require the partial transfer of weights (but we can stream by height). We can CIN-split when we need to split by input channel, this is the last resort because it is also the most expensive (accumulation of the outputs). Once we have the subvolume, we know if input and output tensor need W-split or CIN-split. This in general, means that the tensor cannot fit completely in Mem-tile and part of it will be moved. Or the computation will need to read from mem-tile the same inputs or weights multiple times.

At this stage, the subvolume describes the smallest shape of the weights we need to manage. We compress the weight accordingly to such a subvolume so that any data movement will always be a multiple of the subvolume and can be a single load. Such a compress data will have the same properties whether it is sparse or dense. Of course, sparse data is smaller and we compute fewer operations.

4.3 Schedule and memory allocation

This is a two level memory hierarchy. During the scheduling of each layer, we evaluate what can fit in mem tile. Here, activation and weight tensors share the space. It means that an input tensor is distributed among the memtiles identified by one starting address and a final address and weights are distributed similarly with a start and a final address. At each step the memory allocation will check if we can allocate a tensor in memtile. If we cannot, we evict all tensors into DDR and then split/time the computation.

At the end of this stage, every tensor will be associated to an address in memtile or DDR (or both). If there are only DDR addresses, the compiler will take the basic computation and, by heuristics, will split the computation (thus the output tensor) by width, output channel, height, and input channel (non necessarily in this order). Every computation will have the DDR to and from MemTile and its data movements for the first two levels of the memory hierarchy. The heuristics have a simple objective: find the largest problem fitting the memory level, with the assumption that the output tensor will be build by row and streaming using ping/pong double buffering (considering padding, strides and kernels sizes).

We use an implementation that takes a recursive and iterative approach in tiling [8]: For clarity, $\sum$ is a sum or reduction and $\dot{\sum}$ is a parallel loop and a W-split can be written as

$$Y = Conv(\boldsymbol{X}, \boldsymbol{W}) = \dot{\sum}_w Conv(\boldsymbol{X}_w, \boldsymbol{W}) \tag{7}$$

The split is pre-computed as function of the foot print, before and after each convolution there will be an explicit data movement. At this stage each input, output, and weights will have an address associated with each sub-computation. Consider a subcomputation as data movement, computation, and data movement, and data movements are optional or decorators. Then the code generation of each $Conv(\boldsymbol{X}_w, \boldsymbol{W})$ is independent and, recursively and as needed, there will be specific splits of the computation accordingly. This is a tree (i.e, a root with k children). Sub convolutions so determined may have different memory requirements, we try as much as possible to obtain symmetric computations (some columns will have to compute slightly larger than needed tensors). If the convolution has strides and large kernel, each sub-convolution may have overlap data but defined addresses and data movement if necessary. For example, computing the output by rows and the weights are reused.

$$\forall w,\ Conv(\boldsymbol{X}_w, \boldsymbol{W}) \rightarrow \dot{\sum}_h Conv(\boldsymbol{X}_{w,h}, \boldsymbol{W}) \tag{8}$$

4.4 Code Generation

The compiler recursively creates a list of operations smaller and smaller that can actually be executed Mem-Tile to Mem-Tile. In practice, there is a further decomposition using only AIE cores but it is completely determined by the subvolume computation as we explained previously. Then, we can explicitly determine the data movements from mem-tile to core and back. Here, we shall show how we produce the execution time estimates as in Figure **??**. Do not adjust the paper or do, on the vertical line there is time in seconds and it evolves by going up (if you flip, it goes from left to right); on the horizontal line, there are the instructions separated to create a sequence of time series. We shall explain further in the next sections.

An important feature of the current system is the concept of **iteration**: Using locks and chaining them (locks like in semaphores): We can write a single instruction from the prospective of a single core (as a SIMD instruction) but driving all cores at once (ASM-like code) for multiple iterations:

```
LOADFM Lock k_0 mem-tile addr core addr iter i
CONV iteration      i
WRITEFM Lock k_1 mem-tile addr core addr iter i
```

There is an implicit lock (say k_x) that is used for the pong and the system cycles in between locks (k_x and k_0). These three operations will be execute a number of iterations i and using a ping/pong they will load different slices of data and compute different slices of data. The key ingredient for this to work is the volumes of input and output tensors are all the same. In our environment padding is common and we can manage by a custom load (from memtile to core) and this requires a custom load that will not be repeated

```
## Head top padding < 50 us First comp block
LOADFM Lock k_0 mem-tile addr_0 core addr iter 1
CONV iteration 1
WRITEFM Lock k_1 mem-tile addr_1 core addr iter 1
## Body iteration > 50 us < 150 us
## k_0 -> k_2 -> k_4 Lock Chain
LOADFM Lock k_2 mem-tile addr_2 core addr iter 9
CONV iteration 7
WRITEFM Lock k_3 mem-tile addr_3 core addr iter 9
## tail  > 150 us Last computation block
LOADFM Lock k_4 mem-tile addr_4 core addr iter 1
CONV iteration 1
WRITEFM Lock k_5 mem-tile addr_5 core addr iter 1
```

See Figure **??** for how this code will play out in practice. At this stage we have all the information we need. Per layer, the code generation is a two pass process: first, we generate code for the all load/store and then we combine them into chain having dependency so that to be logically correct and as fast as possible. Take the code above and we introduce a chain information.

4.5 Time Estimation

At this stage, we need to explain how we can capture the execution time and visualize it as in Figure **??**. Do not adjust the paper or

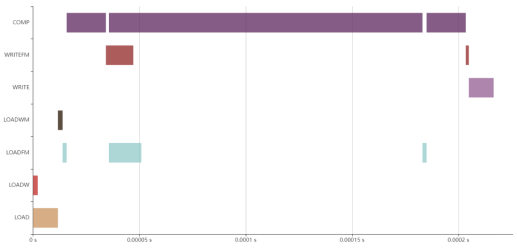

Figure 3: Resnet single convolution with padding for 4x4: legend AIE: LOAD Activation DDR to Mem, LOADW weights DDR to Mem, LOADFM Activation Mem to AIE2 cores, LOADWM weights Mem to AIE2, WRITE Mem to DDR, WRITEFM AIE2 to Mem, COMP Computation

do, on the vertical line there is time in seconds and it evolves by going up (if you flip, it goes from left to right); on the horizontal line, there are the instructions separated to create a sequence of time series. We shall explain further in the next sections.

An important feature of the current system is the concept of **iteration**: Using locks and chaining them (locks like in semaphores): We can write a single instruction from the prospective of a single core (as a SIMD instruction) but driving all cores at once (ASM-like code) for multiple iterations:

```
LOADFM Lock k_0 mem-tile addr core addr iter i
CONV iteration      i
WRITEFM Lock k_1 mem-tile addr core addr iter i
```

There is an implicit lock (say k_x) that is used for the pong and the system cycles in between locks (k_x and k_0). These three operations will be execute a number of iterations i and using a ping/pong they will load different slices of data and compute different slices of data. The key ingredient for this to work is the volumes of input and output tensors are all the same. In our environment padding is common and we can manage by a custom load (from memtile to core) and this requires a custom load that will not be repeated

```
## Head top padding < 50 us First comp block
LOADFM Lock k_0 mem-tile addr_0 core addr iter 1
CONV iteration 1
WRITEFM Lock k_1 mem-tile addr_1 core addr iter 1
## Body iteration > 50 us < 150 us
## k_0 -> k_2 -> k_4 Lock Chain
LOADFM Lock k_2 mem-tile addr_2 core addr iter 9
CONV iteration 7
WRITEFM Lock k_3 mem-tile addr_3 core addr iter 9
## tail  > 150 us Last computation block
LOADFM Lock k_4 mem-tile addr_4 core addr iter 1
CONV iteration 1
WRITEFM Lock k_5 mem-tile addr_5 core addr iter 1
```

See Figure **??** for how this code will play out in practice. At this stage we have all the information we need. Per layer, the code generation is a two pass process: first, we generate code for the all load/store and then we combine them into chain having dependency so that to be logically correct and as fast as possible. Take the code above and we introduce a chain information.

4.6 Time Estimation

At this stage, we need to explain how we can capture the execution time and visualize it as in Figure **??**.

4.7 Time estimates for DDR to Mem-Tile

We have two communication types: activations and weights. Per mem-tile there are two dedicated channels.

- If we share activations and weights in the same mem-tile, we can use one channel for activations and one for weights. Thus the loads from DDR to MEM-tile (LOAD and LOADW) are parallel with a bandwidth of 4GB/s. Writes from mem-tile to DDR (WRITE) can use both channels (8GB/s). We shall expand this.
- If activations and weights go to different mem-tiles (2 and 2), each load is parallel and 8GB/s. Writes are identical.

Although, mem-tile are circular buffers to improve the memory allocation, the classic streaming is not really applicable for two main reasons: First, when we share weights and they must stay still to be reused, if weights are not shared then in general inputs and outputs have different rates of consumption and thus eventually one will overwrite the other. Ping/pong techniques will decrease the space available but it will hide latency: at this time we try to increase the reuse.

4.8 Time estimates for Mem-Tile to AIEs

The Memtile connections with AIE cores can be designed differently. We assume here a few channels with again 4GB/s bandwidth. One memetile can broad cast inputs to a cores column (and to the nearest neighbor). These communications are for activations (LOADFM). One Memtile can broadcast to rows of cores (or the nearest neighbor), these are for weights (LOADWM). We assume that the column and row communications are parallel and each memtile core connection is parallel.

Every communication with iteration 1 is synchronous and binding: that is sequential, the load, convolution, and store is one after the other and every core is independent. For synchronization and for bookkeeping, we assume that AIE2 weights communication (from memtile) to core are synchronous and halting.

Every communication with iteration larger than one, we assume that load, computation (COMP), and store (WRITEFM) are executed in parallel and the overall execution time is the maximum of the estimated time multiplied by the number of iterations.

We estimate the execution time of a subvolume by the number of operations divided by the maximum number of operations per cycle which is in our scenario: $4 \times 8 \times 8 = 256$ operations per cycle and 1GHz frequency. This is obviously a very optimistic validation and not only for convolutions (asking for a ratio of 1 in efficiency) but also for operations like average pools and element wise additions. The execution time is a feature of the analysis but for us the estimate of the communications is more compelling and we can easily mute the computation contribution.

We do not account the wait and synchronization which are necessary to reprogram the fabric. These are very expensive running on a few milliseconds.

During and especially once we generated the code for all data movements, data volumes and destinations, the trigger of the computation, and the sub volume per computation we can estimate quite accurately the execution time for each operations and account their parallel execution.

4.9 Convolution example

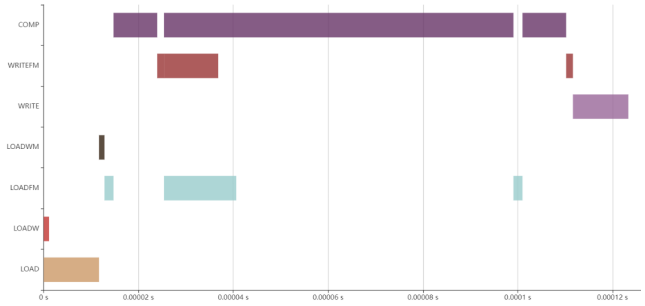

Figure 4: Resnet single convolution with padding and sparsity for 4x4 AIE

Here and in Figure **??** and **??**, we give a full convolution example with and without sparsity and with padding. In this way we can explain how the time is really estimated and also how the hardware works in principle.

It is clear from the Figures above that there are three computations (COMP). We load the weight and activations once in memtiles. There are actually one load per memtile for a total of 4 loads per activations and 4 loads for weights. Because each load is to a different memtile, they are parallel. The activation and weights communications are using two different channels and then are in parallel with 4GB/s bandwidth.

There is a single load of the weights from mem-tiles to each core. This is done once and it is blocking (LOADWM) but it can be as easily made non blocking and parallel to the activations. There is a computation using padding (top-padding) and you can see the sequential execution of load to cores (LOADFM), computation (COMP), and write to memtile (WRITEFM). This computation has iteration 1. There are 9 iterations for three instructions: we can see the load, the computation, and write in parallel. See for the dense convolution in Figure **??** in the time interval between 20μs and 180μs. This is obviously a simplification; there will be a little load poking out at the beginning and a writing poking out at the end. Then we conclude with the final computation with padding at the bottom of the computation.

In this particular scenario, the computation dominates the execution time and compression basically cut the execution time by half: from 200 μs to 130 μs. There are convolutions in Resnet that realize up to 2× performance but also there are convolutions that are dominated by the read or by the writing, and where sparsity help only in space saving.

5 RESULTS

In this section, we present the performance of sparsity applied to all the convolutions (except the first one) for Resnet 50 Figure **??** and **??**, and Inception V3 Figure **??**.

When we generate the code for each instruction, we compute the execution time. So if we inspect the assembly code we will find time information in the context whether or not each instruction contribute directly. For data movement to and from DDR and memtile, we reduce the contribution (sum directly). There is no streaming or communication overlapping, thus the sum.

Table 2: Execution Time estimates

AIE2	Model	Dense sec	Sparse sec
2x2	Resnet	2.492347e-02	1.582626e-02
3x3		1.269543e-02	8.661490e-03
4x4		1.077318e-02	7.064918e-03
5x5		failed	4.303485e-03
6x6		5.712521e-03	4.490127e-03
7x7		4.205991e-03	3.212234e-03
8x8		6.376768e-03	4.602027e-03
2x2	IncV3	4.283837e-02	2.440544e-02
3x3		2.386600e-02	1.422390e-02
4x4		1.740967e-02	1.012540e-02
5x5		9.690552e-03	failed
6x6		1.063962e-02	6.439692e-03
7x7		8.727651e-03	failed
8x8		9.093276e-03	5.666152e-03
2x2	VGG16	4.476212e-02	2.608593e-02
3x3		failed	1.002015e-02
4x4		1.371000e-02	8.852128e-03
5x5		failed	4.336479e-03
6x6		failed	5.770197e-03
7x7		7.455440e-03	5.288551e-03
8x8		9.203393e-03	6.502333e-03

For mem-tile to and from core communications and core computations, we create a time series. We explain in the previous section how we account for the execution time for instruction with and without iterations. All of this will be an attribute of the layer (node in the graph computation). To create a complete time estimate, we just need to take the schedule of the computation and the graph, visit each node accordingly to the schedule, write a *json* file describing the time series, then by *javascript* we can visualize the time series using a browser. The Figures in this paper are generated directly.

For a 4×4 AIE set up, Resnet 50 fits in memtile from the beginning to the last operation (beside the first convolution and this is by costruction). The estimates advantage by sparsity is almost completely achieved. See Figure **??** and **??**.

For the estimate of Inception V3, we use a 6×6 AIE set up. This allows to have a limited spill into DDR and thus the sparsity delivery once again an overall 2× speed up. See Figure **??**. Inception V3 is mostly convolutions of different sizes.

6 CONCLUSIONS AND CONTEXT

This is a multifaceted problem and we present a complete solution from training techniques, compilers, code generation, HW definition, and time estimations.

This could be seen as a quantization and sparsification problem: How to reduce the footprint of a CNN network. There are post training techniques that are targeting quantization and unstructured

sparsity [11] and all the references within. We need to be more aggressive and training for it (as starting point [12]), our sparsity is not really a property of the model and software can describe it and the hardware can take advantage.

This work stemmed from a collaboration between Xilinx and Numenta and the idea set forward by the authors in [2], where the model are too dense and there is a lot we can do about it.

The difficulty of generate code for complex architectures can be described and solved in different ways. There are recent attempts of introducing SW/HW solution for spatial processors like ours [4, 13, 17].

REFERENCES

[1] [n.d.]. *TensorFlow.* https://www.tensorflow.org/

[2] Subutai Ahmad and Luiz Scheinkman. 2019. How Can We Be So Dense? The Benefits of Using Highly Sparse Representations. arXiv:cs.LG/1903.11257

[3] AMD. 2020. rocSPARSE. https://rocsparse.readthedocs.io/en/master/.

[4] Jingwei Cai, Yuchen Wei, Zuotong Wu, Sen Peng, and Kaisheng Ma. 2023. Inter-layer Scheduling Space Definition and Exploration for Tiled Accelerators. *Proceedings of the 50th Annual International Symposium on Computer Architecture* (2023). https://api.semanticscholar.org/CorpusID:259177591

[5] François Chollet et al. 2015. Keras. https://keras.io.

[6] Paolo D'Alberto. 2013. https://github.com/paolodalberto/SparseFastMM/.

[7] Paolo D'Alberto. 2020. Block Sparsity and Training. https://github.com/paolodalberto/BlockSparsityyTraning.

[8] Paolo D'Alberto, Jiangsha Ma, Jintao Li, Yiming Hu, Manasa Bollavaram, and Shaoxia Fang. 2021. DPUV3INT8: A Compiler View to programmable FPGA Inference Engines. *CoRR* abs/2110.04327 (2021). arXiv:2110.04327 https://arxiv.org/abs/2110.04327

[9] Paolo D'Alberto, Victor Wu, Aaron Ng, Rahul Nimaiyar, Elliott Delaye, and Ashish Sirasao. 2022. XDNN: Inference for Deep Convolutional Neural Networks. *ACM Trans. Reconfigurable Technol. Syst.* 15, 2, Article 18 (jan 2022), 29 pages. https://doi.org/10.1145/3473334

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[11] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. arXiv:cs.LG/2210.17323

[12] Benjamin Hawks, Javier M. Duarte, Nicholas J. Fraser, Alessandro Pappalardo, Nhan Tran, and Yaman Umuroglu. 2021. Ps and Qs: Quantization-aware pruning for efficient low latency neural network inference. *CoRR* abs/2102.11289 (2021). arXiv:2102.11289 https://arxiv.org/abs/2102.11289

[13] Qijing Huang, Minwoo Kang, Grace Dinh, Thomas Norell, Aravind Kalaiah, James Demmel, John Wawrzynek, and Yakun Sophia Shao. 2021. CoSA: Scheduling by Constrained Optimization for Spatial Accelerators. *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)* (2021), 554–566. https://api.semanticscholar.org/CorpusID:233740109

[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, Florida, USA) *(MM '14)*. Association for Computing Machinery, New York, NY, USA, 675–678. https://doi.org/10.1145/2647868.2654889

[15] NVIDIA. 2020. cuSPARSE. https://developer.nvidia.com/cusparse.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8026–8037. http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[17] Enrico Russo, Maurizio Palesi, Giuseppe Ascia, Davide Patti, Salvatore Monteleone, and Vincenzo Catania. 2023. Memory-Aware DNN Algorithm-Hardware Mapping via Integer Linear Programming. *Proceedings of the 20th ACM International Conference on Computing Frontiers* (2023). https://api.semanticscholar.org/CorpusID:260442361

[18] Zhewei Yao, Amir Gholami, Sheng Shen, Kurt Keutzer, and Michael W Mahoney. 2021. ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning. *AAAI (Accepted)* (2021).

[19] Shixing Yu, Zhewei Yao, Amir Gholami, Zhen Dong, Michael W. Mahoney, and Kurt Keutzer. 2021. Hessian-Aware Pruning and Optimal Neural Implant. *CoRR* abs/2101.08940 (2021). arXiv:2101.08940 https://arxiv.org/abs/2101.08940

[20] Ben Zandonati, Adrian Alan Pol, Maurizio Pierini, Olya Sirkin, and Tal Kopetz. 2022. FIT: A Metric for Model Sensitivity. arXiv:cs.LG/2210.08502

Received TBD; revised TBD; accepted TBD