
FLAASH: Flexible Accelerator Architecture for Sparse High-Order Tensor Contraction

Gabriel Kulp¹ Andrew Ensinger¹ Lizhong Chen¹

Abstract

Tensors play a vital role in machine learning (ML) and often exhibit properties best explored while maintaining high-order. Efficiently performing ML computations requires taking advantage of sparsity, but generalized hardware support is challenging. This paper introduces FLAASH, a flexible and modular accelerator design for sparse tensor contraction that achieves over $25\times$ speedup for a deep learning workload. Our architecture performs sparse high-order tensor contraction by distributing sparse dot products, or portions thereof, to numerous Sparse Dot Product Engines (SDPEs). Memory structure and job distribution can be customized, and we demonstrate a simple approach as a proof of concept. We address the challenges associated with control flow to navigate data structures, high-order representation, and high-sparsity handling. The effectiveness of our approach is demonstrated through various evaluations, showcasing significant speedup as sparsity and order increase.

1. Introduction

Tensors are ubiquitous in machine learning, with deep learning models commonly using high-order tensors to represent data, neural activations, and parameters. Sparse tensor computations can increase the efficiency of processing large-scale data. Sparsity presents an opportunity to compress data in storage and skip superfluous computations. However, software approaches have limited ability to harness the parallel nature of the arithmetic, and hardware support remains challenging.

Enhancing the efficiency of *tensor contraction*, an operation that extends matrix multiplication to higher dimensions, could accelerate various types of machine learning (Dave et al., 2021). Neural network layers can be replaced with a single tensor contraction or tensor decomposition (decomposition requires many contraction operations when optimizing and evaluating) (Kossaifi et al., 2017; 2020). Additionally, Stoudenmire & Schwab (2017) note that tensor

contraction can be used directly as a method of machine learning. Currently, these operations are expensive compared to the standard highly-optimized operations on dense tensors of activations and weights.

There is a pressing need to improve hardware support for high-order tensors with unstructured sparsity. Recent advancements have made progress in addressing challenges in tensor contraction in general, but they do not take on this specific problem. Du et al. (2022) focus only on GPU implementations. Gondimalla et al. (2021) target sparse convolution operations, and Mishra et al. (2021) deal with structured sparsity. It is common for existing work to concentrate on only matrices (Du et al., 2022; Mishra et al., 2021; Zhang et al., 2020; Qin et al., 2021; Wang et al., 2021; Pal et al., 2018), or on sparse-dense operations (Srivastava et al., 2020). Performance improvements from our work will directly enable and enhance the improvements promised in these approaches (e.g., reducing the parameter count by more than 65% while maintaining classifier performance).

Although structured sparsity can be exploited (Mishra et al., 2021), imposing structure requirements can limit the ability to achieve high sparsity levels. A method for computing on and storing unstructured sparse data is essential. Unstructured sparsity is more challenging to implement, since it allows no assumptions, while structured sparsity relies on meeting specific conditions. Typically, sparse tensor hardware focuses on matrices, with higher-order tensors requiring reformatting or posing difficulties in memory access patterns and caching. Arbitrary-order hardware therefore offers the greatest flexibility. These requirements are met by (Hegde et al., 2019), though our approach offers greater flexibility by using modularity to easily support alternative tensor storage formats, scheduling strategies, and caching behavior.

To address these challenges, we introduce FLAASH (FLexible Accelerator Architecture for Sparse High-order tensor contraction), a groundbreaking approach to sparse tensor contraction architecture. Our results demonstrate the significant potential of FLAASH. We evaluate its performance in two categories of benchmarks: synthetic and deep learning.

¹School of EECS, Oregon State University, Corvallis, OR. Correspondence to: Gabriel Kulp <kulpga@oregonstate.edu>, Lizhong Chen <chenliz@oregonstate.edu>.

The main contributions of this paper are:

1. Proposing a novel architecture for computing tensor contractions with modular flexibility in both the arithmetic decomposition and the data formats.
2. Implementing reference hardware of the architecture to contract sparse tensors in Compressed Sparse Fiber format.
3. Evaluating the hardware implementation, via a conservatively simulated implementation, by conducting experiments against a common data science software package for sparse tensor operations.

2. Background and Related Work

A *tensor* is a multidimensional array, used to represent and manipulate multilinear data. Each dimension of a tensor is called a *mode*, and the number of modes is the tensor’s *order*. A matrix, for example, has order 2 and the two modes are rows and columns. We denote tensors with calligraphic capital letters, as in $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ (a tensor with three modes), and use x_{ijk} to represent the element at coordinates (i, j, k) . A subset of coordinates can also be shown as a list $\{a\}$ for brevity; for example, if $\{a\} = \{i, j\}$, then we can say $x_{i,j,k} = x_{\{a\}k}$.

Tensors have many uses as practical data formats. For example, a gray-scale image is an order-2 tensor, and a color image is order-3. But tensors can be used for more than just storage: they are also the way to arrange operands for computations that, as discussed in Section 1, are useful for machine learning.

One such computation, *tensor contraction*, extends the idea of a matrix multiplication into higher orders. Contraction yields the inner-product of two tensors. In a matrix multiplication, the operation is

$$\mathcal{C}_{ik} = \sum_j \mathcal{A}_{ij} \mathcal{B}_{jk} \quad (1)$$

which is a series of dot products matching up every column of one matrix with every row of the other to form one entry of the result. The result of a matrix multiplication is another matrix, which has order 2. More generally, the order of the result of a contraction is $m + n - 2$ where m and n are the orders of the input tensors. Just like in matrix multiplication, the lengths of each contributing fiber to a tensor contraction must be equal.

Fiber is a generalization of the concept of a row or column. In tensor contraction, the *fiber* length of one operand’s contraction mode must match the fiber length of the other. Therefore, more generally, $\mathcal{A} \in \mathbb{R}^{I \times J \times K}$ and $\mathcal{B} \in \mathbb{R}^{X \times Y}$ can contract along dimensions J and Y only if J and

Y have the same length. This operation yields a result $\mathcal{C} \in \mathbb{R}^{I \times K \times X}$ where the shapes of $\mathcal{A}$ and $\mathcal{B}$ are initially concatenated (I, J, K, X, Y) and then the contracting modes are removed, leaving modes I, K, X . The modes that are not contracted are the ones traversed to find fibers to match up for dot products. These so-called *free modes* are the same modes represented in $\mathcal{C}$: I, K, X in this case. For clarity, we will only contract along the last mode in this example, so the contraction operation can be written

$$\mathcal{C}_{\{a\}\{b\}} = \sum_i \mathcal{A}_{\{a\}i} \mathcal{B}_{\{b\}i} \quad (2)$$

where $\{a\}$ and $\{b\}$ index the free modes and i indexes the contraction mode.

2.1. Sparsity

Sparse tensors have a high proportion of entries that are zeros, with only some small number of non-zero values. When expressed as a percentage, sparsity is the fraction of the total volume (product of all mode lengths) occupied by nonzero values. The number of non-zero values is often abbreviated as NNZ. These zeros obviate the multiplications and accumulations internal to the contraction operation, meaning that sparsity-unaware storage and compute techniques will “waste time” processing zeros that could be skipped. This goal is at the heart of sparsity optimizations: efficiently skipping the zeros in storage and compute.

Sparsity commonly arises in data used for machine learning for many reasons. Sparse matrices may represent momentary observations in discrete time, such as readings from a photodetector that usually does not detect any photons. Sparsity often arises in the storage of graphs using the adjacency matrix format, where each coordinate in a matrix represents a connection between nodes referenced by the row and column. Since many graphs are far from fully-connected, this leads to many zeros representing no connection between those nodes. Also, sparsity is often desirable in neural network weights (and therefore activations) themselves to reduce storage size or inference FLOP count (assuming the inference procedure is sparsity-aware)(Dave et al., 2021).

In unstructured sparsity, there is potential for significant imbalance in the number of nonzero elements in each fiber. One job may require iterating through 10000 nonzero elements while the next job has only two. This puts strain on the scheduling system, particularly if it cannot predict that these scheduled jobs will take very different amounts of time to complete. This challenge is particularly pressing for GPU-like architectures with a shared instruction pointer across many processing elements since the job that completes early will keep the processing element occupied until the longest-running job in the work group completes. Our architecture instead allows each processing element to

operate independently without adding complexity.

Sparse tensors can be stored in memory and on disk in a variety of formats. *Compressed sparse fiber (CSF)* format is common because it is an extension of the CSR (row) and CSC (column) formats often used for sparse matrices. Storing in CSF format requires performing the following procedure. First, a mode is chosen, and all fibers that go in that direction are enumerated. Next, for each fiber, coordinate-entry pairs are stored in indexed order with all zero elements omitted. Then, these coordinate-entry pairs are stored end-to-end in some way that enables later retrieval of a fiber. Often, the approach is to store the start and end pointers for each fiber in an order that makes traversal easy.

While these patterns are easy for a CPU to parse, a CPU is insufficiently parallel and leaves performance on the table. This means we need a custom architecture to traverse these structures efficiently. We implement hardware traversal of compressed data structures to skip element-wise zero checks or seeking: all pointers are calculated from tensor metadata while tensor in-memory size is dominated by NNZ.

2.2. Prior work

Prior work includes *software* approaches for sparse tensor contractions and also includes hardware accelerators for sparse *matrix* operations. We believe that our *modular hardware* accelerator for unstructured high-order sparse *tensor* contractions is a novel contribution. Our work is similar to (Hegde et al., 2019), but our approach places a strong focus on the modularity of the architecture, leaving the door open to alternative designs of each component while retaining the overall architecture structure. We made decisions to use particular tensor storage formats, scheduling strategies, and memory setups, since concrete options must be chosen to perform simulations and benchmarks, but in this paper we present a general flexible and modular framework for the construction of sparse tensor contraction accelerator architectures, while (Hegde et al., 2019) presents only a single design.

There is disagreement about the optimal choice of data structure for accessing and storing operands, intermediate results, and the final result of contraction. In particular, many papers on the subject discuss the use of the compressed sparse row/column/fiber (CSR, CSC, CSF) format (Du et al., 2022; Tillinghast & Zhe, 2021; Kjolstad et al., 2017) and a coordinate-keyed dictionary format (Liu et al., 2021b;a).

Wang et al. (2021) and Zhang et al. (2020) explicitly work down to the hardware on sparse tensors, but they only consider matrices and outer-products (recall that contraction is an inner-product). Qin et al. (2021) discuss hardware design of format converters for working with sparse matrices during multiplication, which play a similar role to the “tensor

memory” portion of FLAASH, discussed in Section 3. Their work could be adapted to serve as an implementation of one modular unit in our full architecture.

Prior work also shows that sparse tensor contraction can play an important role in machine learning. Tensor contraction is explored in deep learning as a replacement for fully-connected layers (Kossaifi et al., 2017; Stoudenmire & Schwab, 2017), and as novel layers (Kossaifi et al., 2020). There is recent work on noticing and promoting sparsity in neural network activation tensors (Andriushchenko et al., 2022; Li et al., 2022) and in transformer activations.

3. Approach

We present FLAASH: a FLExible Accelerator Architecture for Sparse High-order tensor contraction. This architecture scales well to the available die area, and it efficiently handles sparsity: computation time is dominated by the number of nonzero values rather than the volume of the tensors. The design is modular in that each component communicates through an abstracted interface, allowing support for multiple storage formats, cache-aware job scheduling, and optional scratchpad use.

Our design offers a complex control flow in a highly parallelized implementation: each SDPE follows its own state machine to traverse through the sparse dot product, but is small enough that many fit on the die.

Finally, our architecture achieves workload balance across processing elements (Sparse Dot Product Engines, or SDPEs), facilitated by the central job queue. Even if one job specifies a fiber that is more dense than average, the other jobs are not held back from dispatch or completion. Our architecture instead allows each SDPE to operate independently without adding complexity. Unlike the GPU latency hiding approach of context switching, our design leverages a surplus of processing elements such that even a high idle fraction can saturate the memory interface.

3.1. Overview

FLAASH has four high-level components, shown in Fig. 1: processing elements called *Sparse Dot Product Engines (SDPEs)*, job generator & dispatch, tensor memory, and input/output. The arithmetic approach is to convert the tensor contraction into a list of dot products, then distribute these dot products over the set of SDPEs. Each SDPE is a small, simple element that traverses the tensor data structure to process a single dot product. The SDPE is responsible for gathering all relevant tensor elements, multiplying and accumulating as needed, and requesting that the final accumulator value be written into the result tensor.

The general process this architecture aims to implement is

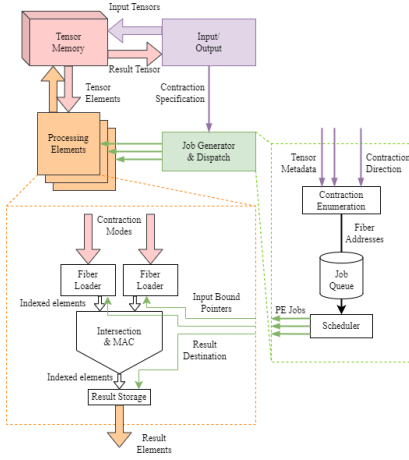


Figure 1. FLAASH architecture overview in the center. Detail views below.

Algorithm 1: Architecture Level Algorithm

Data: Tensor $\mathcal{A}$
Data: Tensor $\mathcal{B}$
Result: Tensor $\mathcal{C}$
 Save entries of $\mathcal{A}$ and $\mathcal{B}$ to Tensor Memory;
 Save pointers of $\mathcal{A}$ and $\mathcal{B}$ to Job Generator;
 Calculate Tensor $\mathcal{C}$ Metadata;
 $\text{Job Count} \leftarrow (\mathcal{A}_{\text{Pointer Count}} - 1) \times (\mathcal{B}_{\text{Pointer Count}} - 1)$
 $\text{Jobs Assigned} \leftarrow 0$ $\text{Jobs Done} \leftarrow 0$
while $\text{Jobs Done} < \text{Job Count}$ **do**
 Wait for a PE to be ready;
 if PE has Result **then**
 if Result Data $\neq 0$ **then**
 Save Result to $\mathcal{C}$ memory;
 Update Pointers and Entry Count;
 end
 Jobs Done $\leftarrow$ Jobs Done + 1;
 end
 if Jobs Assigned $<$ Job Count **then**
 Generate Job fiber bounds;
 Assign Job to Processing Element;
 Jobs Assigned $\leftarrow$ Jobs Assigned + 1;
 end
end
return Pointer to $\mathcal{C}$ in memory;

shown in Algorithm 1. At the top level, the accelerator takes as input two sparse tensors and the modes along which to contract them. The input/output manager handles any direct memory access requirements or buffer copies to ensure that all required data is available to the tensor memory as needed. The job generator then reads the metadata of these tensors to generate a list of dot products that will accomplish the contraction. These dot products (represented as the portions of the input tensors to be read) are then distributed to the SDPEs, which handle the translation from dot product ranges to tensor indices, and request elements from the tensor memory. The tensor memory then services these requests, responding with tensor elements and their index along the fiber.

The scheduler maintains a queue of jobs and distributes them to SDPEs as they finish each dot product job. When all jobs are complete and the last result is written, the accelerator signals that it has finished the contraction and returns a pointer to the resulting tensor data structure.

Algorithm 2: Sparse Dot Product Engine (SDPE)

Data: Tensor $\mathcal{A}$ Fiber Start and End Pointers
Data: Tensor $\mathcal{B}$ Fiber Start and End Pointers
Data: Tensor $\mathcal{C}$ Fiber Entry Destination and Index
 $\mathcal{A}_{\text{Pointer}} \leftarrow$ Tensor $\mathcal{A}$ Fiber Start Pointer
 $\mathcal{A}_{\text{End}} \leftarrow$ Tensor $\mathcal{A}$ Fiber End Pointer
 $\mathcal{B}_{\text{Pointer}} \leftarrow$ Tensor $\mathcal{B}$ Fiber Start Pointer
 $\mathcal{B}_{\text{End}} \leftarrow$ Tensor $\mathcal{B}$ Fiber End Pointer
while $\mathcal{A}_{\text{Pointer}} < \mathcal{A}_{\text{End}}$ and $\mathcal{B}_{\text{Pointer}} < \mathcal{B}_{\text{End}}$ **do**
 Entry $\mathcal{A} \leftarrow \mathcal{A}_{\text{Memory}}[\mathcal{A}_{\text{Pointer}}]$
 Entry $\mathcal{B} \leftarrow \mathcal{B}_{\text{Memory}}[\mathcal{B}_{\text{Pointer}}]$ Entry $\mathcal{C}_{\text{data}} \leftarrow 0$
 if Entry $\mathcal{A}_{\text{index}} == \text{Entry}\mathcal{B}_{\text{index}}$ **then**
 Entry $\mathcal{C}_{\text{data}} \leftarrow$
 Entry $\mathcal{A}_{\text{data}} \cdot \text{Entry}\mathcal{B}_{\text{data}}$ $\mathcal{A}_{\text{Pointer}} \leftarrow$
 $\mathcal{A}_{\text{Pointer}} + 1$ $\mathcal{B}_{\text{Pointer}} \leftarrow \mathcal{B}_{\text{Pointer}} + 1$
 end
 else if Entry $\mathcal{A}_{\text{index}} > \text{Entry}\mathcal{B}_{\text{index}}$ **then**
 $\mathcal{B}_{\text{Pointer}} \leftarrow \mathcal{B}_{\text{Pointer}} + 1$;
 end
 else
 $\mathcal{A}_{\text{Pointer}} \leftarrow \mathcal{A}_{\text{Pointer}} + 1$;
 end
end
 Write $\mathcal{C}$ Fiber Entry to Destination;

3.2. Sparse Dot Product Engines

The SDPE design is optimized to be small so that many SDPEs can be tiled in limited die area, and to scale the design to fit the needs of the larger processor design. One SDPE is comprised of two fiber loader units, an intersection & MAC unit, a result storage unit, and a local job queue.

An in-depth view of an SDPE can be seen in Figure 1 and the algorithm is shown in Algorithm 2. The *fiber loader units* maintain a local FIFO (first-in-first-out queue) of sequential nonzero tensor elements, represented as pairs of coordinate and value. Each FIFO represents a portion of one fiber to contract, with the fiber getting progressively fetched and loaded at the top, and entries getting pulled from the bottom to accumulate the dot product. The fiber loader units are responsible for communicating with the tensor memory to request chunks of the tensor data structure; they traverse their local view of the data structure to know what to fetch next as the dot product progresses.

The *intersection & MAC unit* pulls from these FIFOs to search for locations where nonzero entries of each fiber align. When the index within the fiber is the same for both inputs, a collision occurs. This triggers a multiply-accumulate (MAC) inside the intersection unit, adding to its internal accumulator. If the index from fiber loader unit A is greater than that of B, then the intersection unit discards the current element from the B fiber and pops the next from that FIFO. After advancing one fiber, the index comparison is repeated.

The *local job queue* serves as a small buffer, giving the central job queue more freedom to distribute one job per cycle, even if multiple SDPEs are able to start work in the same cycle. With a local job queue, each SDPE can receive its next job and start immediately when it finishes the current one.

Each fiber loader unit also presents a flag to the intersection unit indicating that it has finished adding new pairs to its FIFO. When this flag is raised and the queue is empty, the job is done. This triggers sending the accumulator value to the *storage unit*, which queues until the tensor memory is ready to write the result. This final queue (another FIFO) allows for greater flexibility in the operation of the tensor memory: For example, the job generator could recognize that there is enough room to store all results in SDPE result queues, and signal that the tensor memory should prioritize all read operations before accepting write requests. The memory queue also helps address load imbalance issues, since an SDPE can begin working on its next job immediately, even when the execution time of a job is very short from having few nonzero elements, or even an unexpectedly all-zero fiber.

Note that the interface between a fiber loader unit and memory only returns nonzero tensor elements. This means that the intersection unit is able to cheaply skip ahead while performing the sparse dot product, since seeking does not involve fetching zeros. Even if the tensor memory uses a data format that requires querying the operand tensors repeatedly to find nonzero elements, this inefficiency is hidden from the SDPE execution model, freeing the implementation

Table 1. Generated Contractions

Job #	Tensor $\mathcal{A}$ Fiber	Tensor $\mathcal{B}$ Fiber
0	1	1
1	1	2
2	1	3
3	2	1
4	2	2
5	2	3

of the tensor memory to privately maintain any metadata useful to efficiently traversing the data structure.

3.3. Job Generation and Dispatch

A *job* in this context is defined as a specific dot product represented by the information that the SDPE needs to request input elements and store the result. In mathematical terms, each SDPE calculates

$$\mathcal{C}_{\{a\}\{b\}} = \sum_{i=n}^m \mathcal{A}_{\{a\}i} \mathcal{B}_{\{b\}i} \quad (3)$$

where the job specifies m and n (the limits of a partial or entire dot product), $\{a\}$ and $\{b\}$ (the coordinates along the free modes), and the destination index (where to store the result). To create a list of these specifications, the job generator iterates through all possible values $\{a\}$ and $\{b\}$; it derives the destination index for $\mathcal{C}$ by concatenating these lists and removing the contraction nodes. This is accomplished in a control flow similar to nested `for` loops: the inner loops iterate over $\{a\}$ and the outer loops iterate over $\{b\}$, both skipping the contraction modes of $\mathcal{A}$ and $\mathcal{B}$. Mathematically:

$$\mathcal{A}_{\text{Start}} = \text{Job\#} / (\mathcal{B}_{\text{Pointer Count}} - 1) \quad (4)$$

$$\mathcal{B}_{\text{Start}} = \text{Job\#} \% (\mathcal{B}_{\text{Pointer Count}} - 1) \quad (5)$$

$$\text{Job Count} = (\mathcal{A}_{\text{Pointer Count}} - 1) \times (\mathcal{B}_{\text{Pointer Count}} - 1) \quad (6)$$

Where $\mathcal{A}_{\text{Start}}$ is the index of the pointer pointing to the start of fiber $\{a\}$, and $\mathcal{B}_{\text{Start}}$ is the index of the pointer pointing to the start of fiber $\{b\}$. For example: assume Tensor $\mathcal{A}$ has 2 free modes (3 pointers), and Tensor $\mathcal{B}$ has 3 free modes (4 pointers). Using equations 4 - 6, the following list of contractions shown in table 1 is generated which will cover all possible fibers in the order that they should be added to memory for Tensor $\mathcal{C}$. In this example, the fourth job will contain the pointers pointing to the beginning and end of the second fiber of $\mathcal{A}$, and the pointers pointing to the beginning and end of the first fiber of $\mathcal{B}$. There will be a total of $2 \cdot 3 = 6$ jobs for this contraction.

In our proof-of-concept implementation, the job generator is in charge of translating ranges of tensor indices into pointer boundaries (to mark the start and end of a continuous run

in memory) for $\mathcal{A}_{\{a\}_i} \forall i \in I$. Generated jobs are added to a queue to await dispatch to SDPEs.

When all iteration through the free modes of $\mathcal{A}$ and $\mathcal{B}$ is complete, the job generator waits until the queue is empty, continues to wait until all SDPEs are idle, and then signals that the contraction operation is complete.

Note that dot products can be decomposed into smaller parts. For example:

$$\sum_{i=1}^{100} \mathcal{A}_i \mathcal{B}_i = \sum_{i=1}^{50} \mathcal{A}_i \mathcal{B}_i + \sum_{i=51}^{100} \mathcal{A}_i \mathcal{B}_i. \quad (7)$$

Our architecture could easily be adapted to take advantage of dot product decomposition to optimize for cache residency. This chunking could be implemented in the job generator to maintain simplicity in the SDPEs.

We did not implement this method since our focus was not on optimizing memory structure or cache utilization, but the architecture provides this flexibility. The tensor memory and the job generator must be in agreement about the storage format of the result tensor. Focusing on the common case where the result is more dense than the operands, our simple solution is to store the result tensor in a dense CSF format that does not require any prediction, balancing, or dynamic allocation. We leave it to the driver software to sparsify the result tensor. This design decision allows more freedom in implementing the job generator: it removes write order dependencies that are required for some sparse storage formats, and it also avoids the dynamic requirements of order-independent COO formats (such as trees and hash tables).

3.4. Tensor Memory

The role of the *tensor memory unit* is to interact with system memory on behalf of the SDPEs, performing the translation from tensor operation to memory access. Recall that tensor contraction requires computing on pairs of fibers, where every fiber of the first tensor must be paired with every fiber of the second. As mentioned in the prior section, this carries caching implications. Furthermore, memory may be virtually segmented into read-only operands and a write-only result. There is plenty of flexibility to implement this unit to efficiently handle caching with the unique needs of any choice of tensor data structure. From an integration perspective, it may be helpful to also implement allocation in this unit, and facilitate the import and export of data if a separate memory is used rather than direct memory access (DMA) to the host memory.

The tensor memory performs allocation of tensors for operands and the result by maintaining an internal division between allocated and free memory. The job generator performs the translation from tensor index to memory pointer,

though, which only works because of our choice of data format, which we explicitly chose to lead to the smallest implementation. This means that the SDPEs make memory requests of the tensor memory with direct physical pointers rather than indices. This works well because the CSF format is simple and allows for up-front pointer calculations without dynamic pointer chasing or reallocation. If we had chosen a more complex format like a tree or locality-binned hash table, then the tensor memory would be the only logical place to handle state while traversing the operand structures and assembling the result structure.

This fulfils the adjacency requirement: adjacent elements of a fiber in the contraction mode are also stored physically adjacent in memory. This storage format also allows the tensor memory to preallocate the entire result tensor, which gives the job generator sufficient context to compute the pointer of each result element before the contraction starts. Our implementation preallocates a dense result in CSF format, which can be converted to a sparse representation in one pass. This is suboptimal but simplifies the implementation considerably, and since results tend to be more dense than operands, it is not an invalid assumption.

3.5. Input/Output Interface

We envision this accelerator as a small component of a larger processor design, sharing system memory through a DMA interface. The I/O unit would implement this interface. In a discrete design, this unit would house the PCIe connection to the host system. We implemented this portion of the simulator as part of the testbench, since the design of high-speed IO interfaces is out of scope for this project.

4. Evaluation

4.1. Implementation

We synthesized and simulated a Verilog implementation of our architecture using Xilinx Vivado HLS. Our simulation assumes a conservative 1GHz clock speed, and only times the contraction operation (not copying operands in or the result out of memory). This is a reasonable assumption for a small accelerator on a CPU that has direct memory access to main system DRAM. For comparisons to software alternatives, we evaluate our model against tensor contraction functions found in both the Pytorch¹ and TensorFlow² libraries, which are widely used in the industry for ML applications. Software measurements were performed using an 8-core Intel Core i7-11375H (3.3GHz) with 32GB of system RAM.

¹www.pytorch.org/docs/stable/torch.html

²www.tensorflow.org/resources/libraries-extensions

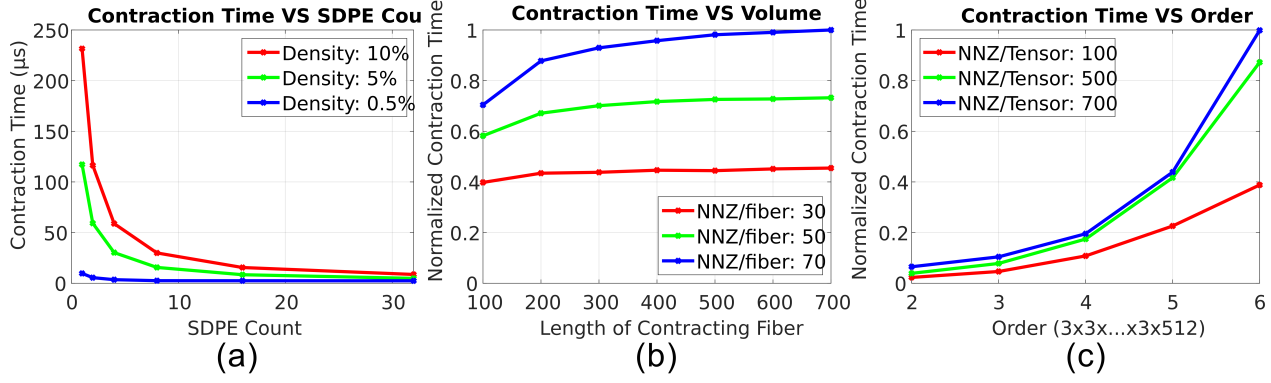


Figure 2. Simulations of Architecture Implemented in Verilog

Using Intel 16nm technology, the relationship between the total area of the architecture as a function of the number of processing elements used is shown in table 2. Eight SDPEs were used for the simulation results found in figures 2b,c and 3a-c.

Table 2. Size of FLAASH Hardware Using Intel 16nm Technology

SDPE Count	Total Area (mm^2)
1	0.692
2	0.703
4	0.721
8	0.750
16	1.938
32	3.129

Tensors are generated randomly, with density as the probability that an individual element will be nonzero. We consider the workloads described by Kossaifi et al. (2017): multi-layer perceptrons from AlexNet and VGG-19, replaced with equivalent tensor contractions. We chose densities to cover a range of empirical results. For example, Sanh et al. (2020) suggest pruned BERT-base encoders achieve around 3% density, while recurrent neural networks from Zhu & Gupta (2017) are below 10%. Andriushchenko et al. (2022) suggest as low as 6.5% density, and Li et al. (2022) suggests 3% to as low as 0.5% density in transformer activations for some layers.

4.2. Synthetic Workload

Fig. 2a demonstrates the relationship between the contraction time and the number of SDPEs. The tensors contracted in this set of simulations had the shape $7 \times 7 \times 512$. It is valuable to note that as the density decreases, the SDPE count becomes less important because the time it takes to complete a single job is faster than the time it takes to read results and assign jobs to all other SDPEs before looping

back. This behavior is a result of our sequential round-robin job distribution strategy. Even at 10% density there is little improvement with adding more than 32 processing elements. The architecture was given 8 SDPEs for the remaining simulations.

In Fig. 2b, we can see that as the Number of Non Zero elements (NNZ) is held constant, there is little difference in contraction time even as the volume of the tensor is increased 7-fold. This is the objective of the architecture: we aim to make the contraction time dependent on the number of non-zero elements in the tensor and not on its dimensions. Only non-zero elements are represented in tensor memory, so the additional volume is only accounted for in the pointers. Note that the contraction time increases proportionally to NNZ in each contracting fiber. The shape of the tensors used in this set of simulations is $5 \times 5 \times n$.

In Fig. 2c, the contraction time is compared to the order of the tensor. In this set of simulations, if the contracting tensor is of order N , the first $N - 1$ dimensions have length 3 and the contraction dimension is of length 512. These tensors are contracted against a 3×512 matrix. Note that the contraction time does increase with order because, as the number of pointers increases, the number of jobs does too, but the contraction time will increase at a rate far lower than the rate the volume is increasing (N^3) if there is a constant number of non-zero elements.

4.3. Machine Learning Workload

Fully Connected Layers (FCLs) typically do not leverage sparsity well and thus consume an unnecessary amount of time because a large number of input nodes are processed to see if they are non-zero. An FCL in a neural network can be replaced by a tensor contraction operation to greatly reduce the number of input parameters Kossaifi et al. (2017). This can be done by using a Tensor Contraction Layer (TCL) which is the process of contracting an input tensor $\mathcal{T}$ of shape $I_1 \times I_2 \times \dots \times I_N$ with a matrix

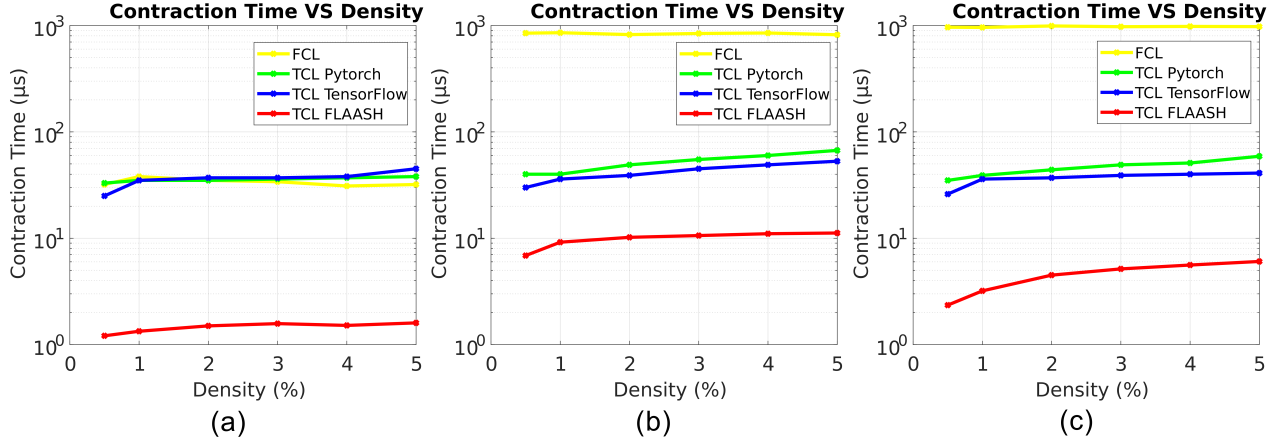


Figure 3. Contraction Times (μs) vs Density (%) (a) $3 \times 3 \times 1024$ tensor contracted with a 3×1024 matrix to get an output tensor of $3 \times 3 \times 3$. (b) $7 \times 7 \times 512$ tensor contracted with a 7×512 matrix to get an output tensor of $7 \times 7 \times 7$. (c) $10 \times 10 \times 100$ tensor contracted with a 10×100 matrix to get an output tensor of $10 \times 10 \times 10$. Matrices involved in contraction each have 50% density.

M of shape $I_N \times R_N$ where $R_N < I_N$ to contract the final dimension to a much smaller size. The matrix M can then be adapted through back propagation to learn dimensional trends. In this section, we compare four schemes. The first is a Fully Connected Layer with $I_1 \cdot I_2 \cdot \dots \cdot I_N$ input nodes and $I_1 \cdot I_2 \cdot \dots \cdot R_N$ output nodes, which will serve as the base case to demonstrate the speedup of using a TCL in ML applications. The second and third schemes are a TCL implemented in software using functions: `torch.sparse.mm()` and `tf.sparse.from_dense()` from two libraries: Pytorch and TensorFlow, respectively. This comparison will demonstrate the CPU performance of executing a TCL and the relative speedup of this method versus an FCL. Finally, the fourth scheme is the implementation of FLAASH which will demonstrate the speedup of using specialized hardware to perform tensor contractions.

The Pytorch and TensorFlow libraries both support dense tensor contraction and sparse matrix contraction, but they do not have functions to perform sparse tensor contraction. An equivalent mathematical operation is to reshape sparse tensors into sparse matrices where the free modes are combined to a single mode. These sparse matrices are then contracted and reshaped to the appropriate dimensions. Note that, for these experiments, only the sparse matrix contraction time is being measured, not the time it takes to reshape.

As can be seen in Figure 3a, when the input tensors have low volume there is no speedup when using a TCL implemented by software, but there is a $23.1\times$ speedup from an average of $33.7\mu\text{s}$ when using an FCL to $1.46\mu\text{s}$ when using the FLAASH accelerator. Speedups for each case in figure 3 can be seen in table 3. There is only $0.375\mu\text{s}$ (30.6%) variation in contraction time as density increases from 0.5% to 5% for the FLAASH accelerator. For the software tensor contractions there is $5\mu\text{s}$ (15%) and $20\mu\text{s}$ (80%) variation for

Pytorch and TensorFlow contraction functions respectively.

Figure 3b shows the contraction times for an input tensor of size $7 \times 7 \times 512$. The contraction time for the fully connected layer has increased to on average $839\mu\text{s}$ with no dependence on density. Tensor Contraction times for Pytorch and TensorFlow implementations stay relatively similar to Fig. 3a at $52\mu\text{s}$ and $42\mu\text{s}$ on average respectively. The contraction times for FLAASH increase to an average of $9.8\mu\text{s}$ due to the increased NNZ and number of free modes.

Finally, in Figure 3c, the contraction times decrease for all TCL implementations because the NNZ is less compared to Fig. 3b but the execution time for the fully connected layer has increased to an average of $975\mu\text{s}$ because of the increased number of output nodes.

Table 3. Avg Speedup of FLAASH in contrast to other methods

Shape	FCL	Pytorch	TensorFlow
(3, 3, 1024)	$23.1\times$	$24.5\times$	$24.8\times$
(7, 7, 512)	$85.3\times$	$5.26\times$	$4.27\times$
(10, 10, 100)	$218\times$	$10.3\times$	$8.16\times$

5. Conclusion and Future Work

Sparsity in tensors is crucial for efficient ML computations, but providing hardware support is challenging. We demonstrate that FLAASH performs well even in high-sparsity and high-order contexts, providing a significant speedup to machine learning workloads. FLAASH opens up possibilities for a variety of future scheduling, cache, and data structure optimizations. Future work could include comparing different data structures and cache management strategies, and exploring opportunities to decompose dot products into

shorter jobs.

References

- Andriushchenko, M., Varre, A., Pillaud-Vivien, L., and Flammarion, N. SGD with large step sizes learns sparse features, October 2022.
- Dave, S., Baghdadi, R., Nowatzki, T., Avancha, S., Shrivastava, A., and Li, B. Hardware Acceleration of Sparse and Irregular Tensor Computations of ML Models: A Survey and Insights. *Proceedings of the IEEE*, 109(10):1706–1752, October 2021. ISSN 1558-2256. doi: 10.1109/JPROC.2021.3098483.
- Du, Z., Guan, Y., Guan, T., Niu, D., Huang, L., Zheng, H., and Xie, Y. OpSparse: A Highly Optimized Framework for Sparse General Matrix Multiplication on GPUs, June 2022.
- Gondimalla, A., Gundabolu, S. C., Vijaykumar, T. N., and Thottethodi, M. Barrier-Free Large-Scale Sparse Tensor Accelerator (BARISTA) For Convolutional Neural Networks, May 2021.
- Hegde, K., Asghari-Moghaddam, H., Pellauer, M., Crago, N., Jaleel, A., Solomonik, E., Emer, J., and Fletcher, C. W. Extensor: An accelerator for sparse tensor algebra. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture, MICRO '22*, pp. 319–333, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369381. doi: 10.1145/3352460.3358275. URL <https://doi.org/10.1145/3352460.3358275>.
- Kjolstad, F., Kamil, S., Chou, S., Lugato, D., and Amarasinghe, S. The tensor algebra compiler. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA): 77:1–77:29, October 2017. doi: 10.1145/3133901.
- Kossaifi, J., Khanna, A., Lipton, Z. C., Furlanello, T., and Anandkumar, A. Tensor contraction layers for parsimonious deep nets. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1940–1946, June 2017.
- Kossaifi, J., Lipton, Z. C., Kolbeinsson, A., Khanna, A., Furlanello, T., and Anandkumar, A. Tensor regression networks. *Journal of Machine Learning Research*, 21(123):1–21, 2020.
- Li, Z., You, C., Bhojanapalli, S., Li, D., Rawat, A. S., Reddi, S. J., Ye, K., Chern, F., Yu, F., Guo, R., and Kumar, S. Large Models are Parsimonious Learners: Activation Sparsity in Trained Transformers, October 2022.
- Liu, J., Li, D., Gioiosa, R., and Li, J. Athena: High-performance sparse tensor contraction sequence on heterogeneous memory. In *Proceedings of the ACM International Conference on Supercomputing, ICS '21*, pp. 190–202, New York, NY, USA, June 2021a. Association for Computing Machinery. ISBN 978-1-4503-8335-6. doi: 10.1145/3447818.3460355.
- Liu, J., Ren, J., Gioiosa, R., Li, D., and Li, J. Sparta: High-performance, element-wise sparse tensor contraction on heterogeneous memory. In *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPoPP '21*, pp. 318–333, New York, NY, USA, February 2021b. Association for Computing Machinery. ISBN 978-1-4503-8294-6. doi: 10.1145/3437801.3441581.
- Mishra, A., Latorre, J. A., Pool, J., Stosic, D., Stosic, D., Venkatesh, G., Yu, C., and Micikevicius, P. Accelerating Sparse Deep Neural Networks, April 2021.
- Pal, S., Beaumont, J., Park, D., Amarnath, A., Feng, S., Chakrabarti, C., Kim, H., Blaauw, D., Mudge, T., and Dreslinski, R. Outerspace: An outer product based sparse matrix multiplication accelerator. In *Proceedings - 24th IEEE International Symposium on High Performance Computer Architecture, HPCA 2018*, Proceedings - International Symposium on High-Performance Computer Architecture, pp. 724–736. IEEE Computer Society, March 2018. doi: 10.1109/HPCA.2018.00067. Publisher Copyright: © 2018 IEEE.; 24th IEEE International Symposium on High Performance Computer Architecture, HPCA 2018 ; Conference date: 24-02-2018 Through 28-02-2018.
- Qin, E., Jeong, G., Won, W., Kao, S.-C., Kwon, H., Srivasan, S., Das, D., Moon, G. E., Rajamanickam, S., and Krishna, T. Extending Sparse Tensor Accelerators to Support Multiple Compression Formats, March 2021.
- Sanh, V., Wolf, T., and Rush, A. Movement Pruning: Adaptive Sparsity by Fine-Tuning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20378–20389. Curran Associates, Inc., 2020.
- Srivastava, N., Jin, H., Smith, S., Rong, H., Albonesi, D., and Zhang, Z. Tensaurus: A versatile accelerator for mixed sparse-dense tensor computations. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 689–702, 2020. doi: 10.1109/HPCA47549.2020.00062.
- Stoudenmire, E. M. and Schwab, D. J. Supervised Learning with Quantum-Inspired Tensor Networks, May 2017.
- Tillinghast, C. and Zhe, S. Nonparametric Decomposition of Sparse Tensors. In *Proceedings of the 38th Inter-*

national Conference on Machine Learning, pp. 10301–10311. PMLR, July 2021.

Wang, Y., Zhang, C., Xie, Z., Guo, C., Liu, Y., and Leng, J. Dual-side Sparse Tensor Core, May 2021.

Zhang, Z., Wang, H., Han, S., and Dally, W. J. SpArch: Efficient Architecture for Sparse Matrix Multiplication. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 261–274, February 2020. doi: 10.1109/HPCA47549.2020.00030.

Zhu, M. and Gupta, S. To prune, or not to prune: Exploring the efficacy of pruning for model compression. *ArXiv*, October 2017.