

Reconfigurable Low-latency Memory System for Sparse Matricized Tensor Times Khatri-Rao Product on FPGA

Sasindu Wijeratne*, Rajgopal Kannan†, Viktor Prasanna*

*Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, USA

†US Army Research Lab, Los Angeles, USA

Email: kangaram@usc.edu, rajgopal.kannan.civ@mail.mil, prasanna@usc.edu

Abstract—Tensor decomposition has become an essential tool in many applications in various domains, including machine learning. Sparse Matricized Tensor Times Khatri-Rao Product (MTTKRP) is one of the most computationally expensive kernels in tensor computations. Despite having significant computational parallelism, MTTKRP is a challenging kernel to optimize due to its irregular memory access characteristics. This paper focuses on a multi-faceted memory system, which explores the spatial and temporal locality of the data structures of MTTKRP. Further, users can reconfigure our design depending on the behavior of the compute units used in the FPGA accelerator. Our system efficiently accesses all the MTTKRP data structures while reducing the total memory access time, using a distributed cache and Direct Memory Access (DMA) subsystem. Moreover, our work improves the memory access time by $3.5\times$ compared with commercial memory controller IPs. Also, our system shows $2\times$ and $1.26\times$ speedups compared with cache-only and DMA-only memory systems, respectively.

Index Terms—MTTKRP, Memory Systems, Shared Memory, FPGA, Tensor Decomposition

I. INTRODUCTION

Tensors are the de-facto representation of high-dimensional data. They have become the center for Machine Learning techniques [1] such as recommender systems [2] [3] and neural networks [4] [5]. Tensors are also used in other domains including network analysis [6], chemistry [7], and signal processing [8].

With the recent surge in Machine Learning, the attention towards tensor decomposition grew exponentially. Canonical polyadic decomposition (CPD) [9] is one of the popular methods for tensor decomposition. CPD approximates the tensor as a sum of rank-one tensors. The alternating least squares method (CP-ALS) is the most common algorithm which use to compute CPD. CP-ALS computes a new factor matrix for each mode in each iteration. Matricized tensor by Khatri-Rao product (MTTKRP), kernel is the common bottleneck in this computation.

Since the sparsity of the real-world tensors [10] [11] is considerably high, specialized hardware accelerators are becoming common means of improving compute efficiency of tensor computations. But external memory access time has become the bottleneck due to irregular data access patterns.

There have been several techniques proposed in the literature to overcome the memory access time while accessing data

with irregular access patterns [12] [13] [14]. Caches [15] are very productive in this regard if the required data fits in the cache and the data has spatial and temporal locality [16]. The ongoing approach for solving long memory access delays is to use onboard Block RAM (BRAM) in the FPGA as a data cache and facilitate data retrieval. An alternative solution is to look at multiple memory requests DMA transfers [14] [17].

In this paper, we propose a memory system to significantly reduce the total memory access time on sparse tensors while performing MTTKRP. We also analyze the memory access patterns of the data structures used in the sparse-MTTKRP operation and suggest the best memory components to use with them. To the best of our knowledge, no prior work has proposed a memory system for sparse MTTKRP compute fabrics while analyzing the memory access patterns of its data structures.

The key contributions of this paper are:

- **Analyzing Memory access patterns of sparse MTTKRP:** We analyze the memory access patterns of different data structures used in sparse MTTKRP computation. Then we propose memory modules (e.g., DMA controller and cache) that can use with the data structures to benefit from their memory access patterns.
- **Reconfigurable memory system:** We propose a Local Memory Block (LMB) based memory system that can reconfigure depending on the targeted compute fabric and data layout used in the accelerator. We also experiment on different types of MTTKRP Processing Elements (PEs) and different memory system configurations, respectively.
- We evaluate the system on a Xilinx Alveo U250 board. It shows $3.5\times$, $2\times$, and $1.26\times$ speedups in memory access time compared with commercial memory controller IPs, cache-only, and DMA-only memory systems.

The rest of the paper is organized as follows, Sections II and III focus on understanding the sparse MTTKRP operation and prior work. Section IV discusses the architecture of the proposed memory design in-depth, while Section V presents the evaluation results. Finally, the paper is concluded in Section VI.

II. BACKGROUND

In this section, we first describe the notations used in the paper. Then we investigate the basics and the usage of MTTKRP kernel.

A. Notation

This paper follows the definitions and symbols used in Tensor Algebra Compiler (TACO) [18]. The only difference is, we use capital calligraphic letters (e.g., $\mathcal{B}$) to represent tensors with N dimensions.

Table I summarizes the symbols commonly used in the paper.

TABLE I: Notation

Notation	Description
$\mathcal{B}$	Tensor
A / C / D	Matrix
$\mathcal{B}_{i,j,k}$	Element at (i, j, k) of $\mathcal{B}$
$A_{i,j}$	Element at index (i, j) of A
nnz	Number of nonzeros in $\mathcal{B}$

B. Use Cases of MTTKRP

MTTKRP is the core computation kernel in the alternating least square (ALS) method, which is a popular method for finding the factor matrices in CPD. The CPD, with R as the decomposition rank, approximates tensor ($\mathcal{B}$) as the sum of R rank-one tensors: $\mathcal{B} \approx \sum_{r=0}^{R-1} a_r \circ d_r \circ c_r$, where $\circ$ denotes the outer product, and a, c, d are the components of the factor matrices A, C, and D. ALS method is a tensor approximation method computes using the above equation. In ALS, one of the factor matrices is updated at a time while fixing the rest. Algorithm 1 shows the steps of CP-ALS algorithm. The MTTKRP operation, which is the focus of this paper, takes place in lines 2, 3, and 4 of Algorithm 1 to compute each of the factor matrices of all modes.

Algorithm 1: CPD-ALS FOR THIRD-ORDER TENSORS

Input: $\mathcal{B} \in \mathbb{R}^{I \times J \times K}$

Output: $A \in \mathbb{R}^{I \times R}$, $C \in \mathbb{R}^{K \times R}$, $D \in \mathbb{R}^{J \times R}$

```

1 while no improvement or maximum iterations reached
  do
2    $A \leftarrow \mathcal{B}_{(1)}(D \odot C)(C^T C * D^T D)$ 
3    $D \leftarrow \mathcal{B}_{(2)}(A \odot C)(C^T C * A^T A)$ 
4    $C \leftarrow \mathcal{B}_{(3)}(D \odot A)(A^T A * D^T D)$ 
5   Normalize columns of A,B,C and store in  $\lambda$ 
6 return A, B, C

```

C. MTTKRP

MTTKRP consists of N-dimensional tensor multiply with N-1 factor matrices. Here, N is the order of the tensor. Mathematically, mode-1 MTTKRP (for three-dimensional tensors) can be expressed as Equation 1. In this equation, A, C, and D are dense matrices while $\mathcal{B}$ is a three-dimensional tensor

(matricized along with the first mode), and $\odot$ denotes the Khatri-Rao product.

$$A = \mathcal{B}_{(1)}(D \odot C) \quad (1)$$

This operation can also express in index notation as Equation 2.

$$A_{i,r} = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} \mathcal{B}_{i,j,k} \cdot D_{j,r} \cdot C_{k,r} \quad (2)$$

In the following sections, we target 3D tensors since they are easy to understand. Even though we focus on 3-dimensional tensors as examples in this paper, we can expand our work into higher dimensions without any extra work.

The sparsity of the tensors creates irregular data access patterns, which makes the external memory access more challenging than the dense MTTKRP. All the real-world tensors are sparse in nature. Algorithm 2 shows the sequential sparse MTTKRP (spMTTKRP) approach for third-order tensors in COO format [19] [20].

Algorithm 2: COO BASED SPMTTKRP FOR THIRD ORDER TENSORS

Input: indI[nnz], indJ[nnz], indK[nnz], vals[nnz], D[J][R], C[K][R]

Output: A[I][R]

```

1 for  $z = 0$  to  $nnz$  do
2    $i = \text{indI}[z]$ 
3    $j = \text{indJ}[z]$ 
4    $k = \text{indK}[z]$ 
5   for  $r = 0$  to  $R$  do
6      $A[i][r] += \text{vals}[z] \cdot D[j][r] \cdot C[k][r]$ 
7 return A

```

We use the term *fibers* throughout this paper. *Fibers* are the building blocks of tensors. Fiber is a one-dimensional fragment of a tensor obtained by fixing all indices but one. Tensor fibers are a higher-order extension of matrix rows and columns. The fibers of a three-dimensional tensor can be represented as $\mathcal{B}_{:,j,k}$, $\mathcal{B}_{i,:,k}$, and $\mathcal{B}_{i,j,:}$. Similarly, for a matrix C, its rows $C_{i,:}$ and columns $C_{:,j}$ are its fibers. In Equation 3 and 4, the term fiber corresponds to the rows of the factor matrices.

III. RELATED WORK

M. Asiatici et al. [21] [22] present miss optimized cache system for FPGA accelerators. The key idea of their work is to increase the reusability to serve multiple requests from the compute fabric on the fly without relying on long-term storage in cache. They also use cuckoo hash table-based MSHR techniques to avoid the memory traffic due to secondary cache misses. In our work, instead of having only a cache-based system, we propose a cache + DMA system to accelerate the external memory accesses. Instead of using MSHR, we introduce an XOR hash-table-based Recent Request Status Holder (RRSH) unit closer to the compute fabric. Unlike MSHR, RRSH takes care of secondary cache misses while

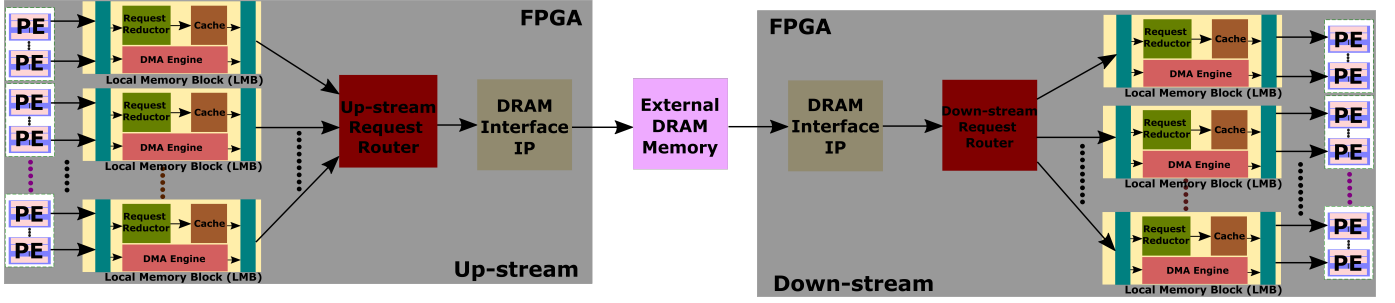


Fig. 1: Overall Architecture

reducing the memory traffic between the cache and compute units. Additionally, we want to emphasize that the focus of our work is entirely on the MTTKRP kernel.

S. Aananthakrishnan et al. [23] have proposed a memory-optimized large-scale graph processor on ASICs. This paper emphasizes the importance of parallel support for bulk transfers and cache transfers for graph workloads. Even though we are focusing on a different domain of algorithms and proposing an entirely different memory system, we still observed the importance of supporting cache and DMA transfers simultaneously.

N. Srivastava et al. [24] [25] have implemented a tensor computation library for FPGAs and CGRAs. These papers describe compute fabrics to accelerate spMTTKRP using novel algorithmic optimizations. In our paper, we explore the memory optimizations we can implement using a multi-faceted approach. The spMTTKRP algorithms describe in [24] [25] can be used as the compute fabrics to our proposed memory system.

IV. MEMORY SYSTEM ARCHITECTURE

We develop a unified memory system to share among multiple PEs that process MTTKRP. The state-of-the-art MTTKRP compute accelerator designs execute either Equation 3 or Equation 4 in their compute fabrics. Here, $fiber_{out}$ represent one or more output fibers of the output matrix, $fiber_j$ and $fiber_k$ represent a single fiber from input matrices, scalar represents a scalar value from high dimensional input tensor, and $\circ$ is corresponding to the Hadamard product of the two input vectors. Even though the internal construction of PEs is different in different state-of-the-art accelerators, they follow the same memory access pattern. The types of memory accesses can be summarized as follows: (a) load the input fibers of the matrices from the external memory, (b) load the scalar of the input tensor from the external memory, (c) store the output fiber to the external memory. In a parallel and pipeline system, (a), (b), and (c) occurs in parallel for multiple fibers. Favorably, our memory system can receive the data from each step simultaneously.

$$fiber_{out} = scalar \cdot \sum_K \sum_J (fiber_k \circ fiber_j) \quad (3)$$

$$fiber_{out} = \sum_K \sum_J fiber_k \circ (scalar \cdot fiber_j) \quad (4)$$

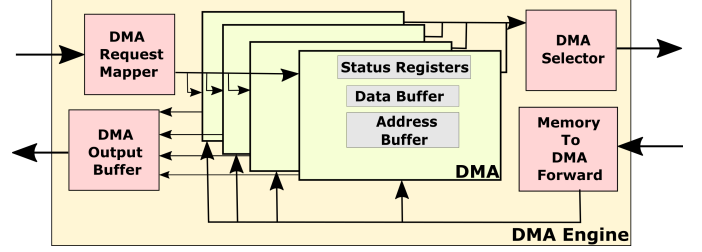


Fig. 2: DMA Engine

Our proposed architecture supports 2 types of memory transfers:

- 1) **cache-line transfers:** Nurtures single memory accesses. Load/store individual requests in minimum latency. The access patterns with high spatial and temporal locality are accessed using cache-line transfers. The element-wise access of scalar values from the input tensor shows spatial and temporal locality. Therefore, we use cache-lines to load the scalars from external memory.
- 2) **DMA transfers:** Supports streaming accesses. Load/store operations on all requested data with minimum latency from memory. Loading and storing matrix fibers is a streaming memory access process. Therefore, we use the DMA engine to transfer matrix fibers between FPGA and external memory.

Figure 1 shows the overall architecture of the proposed memory system. The upstream logic designates the data path from PEs to the external memory, while the downstream logic shows the data path from external memory to the PEs. The overall data path is symmetrical over the external memory.

The Local Memory Blocks (LMBs) are the basic building blocks of our proposed memory system. A LMB has a Request Reducutor, non-blocking cache, and a DMA Engine. Each LMB connects to one or more PEs. The complexity of the connection between PEs and LMB exponentially increases with the number of PEs connected to a single LMB. As shown in Figure 1, we use multiple LMB distributed among all the PEs to avoid such unnecessary complexity. Using more than one LMB does not impact the memory consistency model of the MTTKRP accelerators. The reasons behind maintaining the consistency are: (a) Only the PEs connected to the same LMB update the same output fiber, (b) The input fibers or

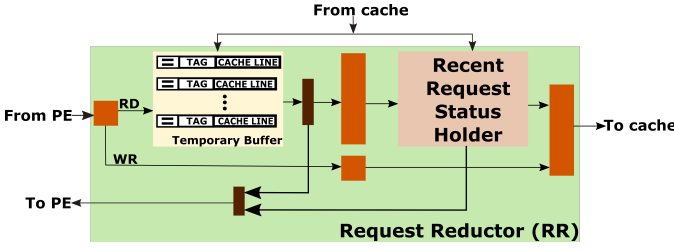


Fig. 3: Internals of Request Reductor

scalars do not update during the same MTTKRP operation.

A. DMA Engine

The DMA engine is in charge of communicating the fibers of the matrices between PEs and the external memory. We store the matrices in row-major order because the MTTKRP algorithm encourages row-wise matrix accesses. The matrices in row-major order make the fiber accesses into a bulk of sequential memory transfers. Figure 2 shows the DMA engine and its internals. It has several DMA buffers inside. Therefore, it can support multiple fiber reads and writes simultaneously. The number of DMA buffers is proportional to the number of PEs connected to the same LMB. A large number of DMA buffers in a LMB can reduce the maximum operating clock frequency due to increased hardware routing complexities.

B. Cache

The cache focuses on satisfying a single memory request with minimum latency. Within the cache engine, we explore the spatial and temporal locality. It loads the scalars of the tensor from the external memory.

Our non-blocking cache uses a 3-stage pipeline to achieve high frequency. We keep the cache-line width similar to the data width of DRAM Interface IP to avoid implementation complexities. Instead of forwarding a single element from the cache to PEs, the cache passes the complete cache-line to the Request Reductor (RR). Then, the RR advances the requested portion to each PE while storing the incoming cache-line inside its temporary buffer.

C. Request Reductor (RR)

RR converts element-wise cache reads to cache-line accesses. As shown in Figure 3, the RR is a 2-stage pipeline. In the first step, a temporary buffer stores the most recent memory reads. It is a CAM-based memory implementation that keeps the most recent external memory reads inside. Since CAMs are hardware expensive, we keep the number of elements in the buffer small.

If requested data is not in the temporary buffer, the read request advances to the Recent Request Status Holder (RRSH). RRSH keeps the status of recently forwarded requests to the cache. If the incoming read request belongs to one of the pending cache-line requests, the PE id and address are kept in the RRSH. When a cache-reply from cache reaches the RRSH, the pending requests corresponding to that cache line

Algorithm 3: PARALLEL-MTTKRP ALGORITHM

```

1 Let  $PE_v$  denote the  $v^{th}$  PE on FPGA,  $0 \leq v < p$ 
2 Let  $Partition_q$  denote the  $q^{th}$  partition,  $0 \leq i < p$ 
Input:  $NNZ[M] = \{indI[M], indJ[M], indK[M],$ 
 $vals[M]\}$ ,  $B[J][R]$ ,  $C[K][R]$ 
Output:  $Y[I][R]$ 
3 while not done do
4   for each partitionq parallel do
5     for  $z = 0$  to  $M/p$  do
6        $i = indI[z]$ 
7        $j = indJ[z]$ 
8        $k = indK[z]$ 
9       if  $current_I \neq indI$  then
10        for  $r = 0$  to  $R$  do
11           $Y[current_I][r] = temp\_Y[r]$ 
12         $current_I = indI$ 
13        for  $r = 0$  to  $R$  do
14           $temp\_Y[r] =$ 
15             $vals[z] \times B[j][r] \times C[k][r]$ 
16        else
17          for  $r = 0$  to  $R$  do
18             $temp\_Y[r] +=$ 
19               $vals[z] \times B[j][r] \times C[k][r]$ 

```

are satisfied by sending the corresponding data elements to the requested PEs. RRRSH helps to convert element-wise cache access to cache-line-wise accesses. It drastically reduces the memory traffic to the cache. RRRSH logic can be implemented using a hash table. In our work, we use XOR-base hash table [26] considering its high throughput and scalability.

1) *XOR-based Hash Tables:* R. Zhang et al. [27] propose XOR-based hash tables that can support multiple parallel pipelines with high throughput. For stall-free execution, our work requires 2 PE versions of the hash table. Each PE connects to the cache interface and PE side separately. The total number of entries in the hash table is proportional to $\frac{\text{total number of entries in the local cache}}{\text{local cache associativity}}$. The width of a table is proportional to $(Tag\ width + Number\ of\ PEs)$ and it connects with $(RR \times number\ of\ data\ elements\ per\ cache - line)$.

D. Request Routers

The request router has 2 responsibilities. They are (a) receive memory requests from different LMB units and forward them to the DRAM interface IP, (b) forward the data coming from external memory to the LMB units.

E. Reconfigurability of the Overall Design

Users can configure our design during the synthesis step. The configuration heavily depends on the characteristics of the compute fabric and data layout. In the current literature, all the FPGA or CGRA based implementations use a variation

TABLE II: Module Configuration and Resource Utilization

Module Name	Specification	Resource Utilization			
		LUT(%)	FF(%)	BRAM(%)	URAM(%)
	Configuration-A				
Cache	Degree of set-associativity = 2 No. of cache-lines = 8192 Cache-line width = 512	1.87	1.24	0.24	1.25
DMA Engine	No. parallel DMA supported = 4 Size of single DMA buffer = 256 B	0.04	0.01	-	0.25
Request Reductor	No. of entries in RRSB = 4096 No. of entries in Tempory Buffer = 8	0.08	0.10	-	1.25
LMB	Includes a cache, a DMA Engine, and a Request Reductor	2.03	1.41	0.24	2.75
Complete System	No. of LMBs = 1	2.25	1.54	0.24	2.75
	Configuration-B				
Cache	Degree of set-associativity = 1 No. of cache-lines = 4096 Cache-line width = 512	0.65	0.64	0.06	0.63
DMA Engine	No. parallel DMA supported = 4 Size of single DMA buffer = 256 B	0.04	0.01	-	0.25
Request Reductor	No. of entries in RRSB = 4096 No. of entries in Tempory Buffer = 8	0.08	0.10	-	1.25
LMB	Includes a cache, a DMA Engine, and a Request Reductor	0.85	0.81	0.06	2.13
Complete System	No. of LMBs = 4	3.61	3.35	0.24	8.52

of COO Format (e.g., CISS) to store the high dimensional tensors. Meanwhile, the dense matrices use a row-major format. We assume all the compute fabrics that use our memory system use the same kind of memory layout. Under these conditions: (a) the sparse tensors use the cache due to the spatial and temporal locality of the data access pattern, (b) The fibers of the matrices are accessed through DMAs due to their high spatial locality and poor temporal locality.

In Section V-C, we dive deeper into configuring the number of LMBs depending on the type of the compute fabric. Our experiments show that the performance improvement due to the total number of DMAs in an LMB saturates after 4 DMAs. Increasing the number of DMAs also negatively impacts the maximum operating frequency due to increased place and route complexity. We further observed that the cache size also influences the maximum operating frequency of the overall design. We keep cache size including, the total number of cache lines and cache line width, as configurable parameters. We maintain the cache line width equal to the memory interface IP's data width since it avoids design complexities of the cache.

V. EVALUATION

A. Experiment Methodology

We implement different configurations of the memory controller on Xilinx Alveo U250 FPGA [28] using Verilog HDL. This device has 1,728 K LUTs, 3,456 K Flip-flops, and 327 MB of URAM memory. Simulations, synthesis, and place-and-route implementations are performed using Xilinx Vivado Design Suite 2020.2. In our work, we focus on accessing a single external DRAM component efficiently. We use the Xilinx memory interface IP [29] to connect our design with the

external memory. For the Alveo U250 board, Xilinx provides Memory Interface IP with 31-bit address width and 512-bit data width (with ECC turned on).

1) *Datasets*: Table III summarizes the characteristics of the synthetic 3D tensors we use in our experiments. Most of the state-of-the-art accelerators use a variation of the COO format to store data. For example, N. Srivastava et al. [24] [25] use Compressed Interleaved Sparse Slice (CISS) format, which is also a variation of COO format. In this section, the tensors use variations of the COO format. Each element in the tensor includes its coordinate vector following the value. The total size of one 3D tensor element is 16 Bytes. We use 32 bits to store each coordinate and value. The dense matrices are stored in row-major order while keeping each element 4 Byte. We set the number of elements in a row of a matrix to 32.

TABLE III: Sparse 3D Tensor Datasets

Tensor	Dimensions	Nonzeros	Density
Synth'01	22K $\times$ 22K $\times$ 23M	28M	2.37E-09
Synth'02	3M $\times$ 2M $\times$ 25M	144M	9.05E-13

B. Baselines

We compare our proposed memory system with 3 approaches.

1) *Memory Controller IP Only*: : Commercial memory interface IP [29] [30] directly connected to a compute fabric of a given accelerator. It is the naïve method of connecting the external memory.

2) *Cache-only*: : Compute fabric of the accelerator connects to the external memory through a cache. This approach is identical to replacing the LMB in our system with the cache.

3) *DMA-only*: All the requests of the accelerator are accessed through a DMA. This approach is identical to replacing the LMB in our system with DMAs. Here, a DMA engine can load/store a single DMA request at a time.

C. Configurations of the Proposed System

There are 2 types of spMTTKRP kernel accelerators based on the communication between external memory and compute fabric.

Type-1: The systolic array-based spMTTKRP compute fabrics [24] [25] in the current literature have a single point of access to the external memory for each type of data structure. After loading the data to the compute fabric, they are shared among PEs using the PEs near the memory system. For instance, [25] use shared Loading and Storing Units for each type of data structure (i.e., Matrix Loading Unit (MLU), Tensor Loading Unit (TLU), and Matrix Store Unit (MSU)) and share the data among PEs using the edge PEs.

Type-2: The compute fabrics with multiple points of access to external memory fall into this category. For instance, Algorithm 3 can be mapped into a compute fabric with multiple PEs with independent memory accesses. Here, the computations executed by each PE are independent of others. The memory system should be sophisticated enough to handle this type of compute fabric.

In the experiments, we use 2 types of configurations depending on the above communication types between external memory and compute fabric.

Configuration-A: For Type-1, having multiple LMBs does not help because their data structures only have a single point of access to the external memory. Therefore, with Type-1 compute fabrics, we use a single Large LMB. Table II shows the resource utilization of each module in the memory system that we used in our experiments.

Configuration-B: For Type-2, multiple LMBs can increase the performance. Their compute fabric has several points of access for each of the data structures. In our experiments, we use 4 LMBs, with each LMB connected to a PE. Table II summarizes the resource utilization of each module in the design we used in our experiments.

D. Performance

As we discussed previously, we ran the experiments on different datasets with different memory system configurations. Figure 4 shows the improvement after using our proposed memory systems over other alternatives. We compute the memory access speedup by comparing the total memory access time with the commercial memory controller IP-only setting. Our proposed system achieved around 3.5x speedup compared to the commercial memory controller IP-only setting. Also, compared with cache-only and DMA-only memory systems, our approach achieved around 2x and 1.26x speedup, respectively.

Our approach always achieves higher speedup compared to the cache-only setting. In MTTKRP, the consecutive secondary cache misses to the same cache line occur while accessing the

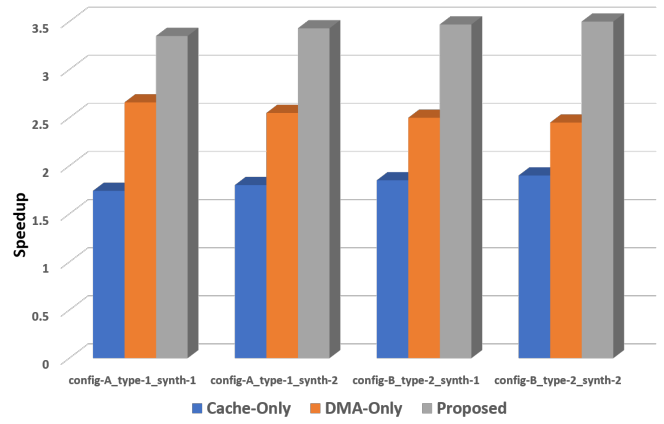


Fig. 4: Speedup with different memory systems over direct connection to commercial Xilinx memory controller IP as the baseline. The naming convention of each category is (<configuration of proposed system> _ <type of the compute fabric> _ <dataset>).

fibers. In modern caches, secondary misses are avoided using Miss Status Holding Registers (MSHR). But conventional MSHR can not handle a large number of secondary cache misses without losing the performance. Further, the memory traffic between the cache and compute fabric can also reduce the performance in the cache-only setting. In our method, we reduce this traffic while avoiding secondary misses using the Request Reductor.

DMAs do not exploit the temporal locality of data. Further, there can be garbage data in DMA transactions when the length of the data requests is shorter than the width of the memory interface IP. Our proposed system outperforms the DMA-only setting simply because it avoids the above issues with DMAs.

VI. CONCLUSION

In this paper, we propose a multifaceted memory system that reduces the total memory access time of MTTKRP on FPGAs. Our unified hardware can be reconfigured in the compile-time, depending on the orientation of the computing units and memory layout of tensors. The scalable Local Memory Blocks can handle the memory access pattern of MTTKRP efficiently while achieving $3.5\times$ speedup compared to the commercial memory controller IPs. Also, our system shows $2\times$ and $1.26\times$ speedups compared with cache-only and DMA-only memory systems, respectively.

VII. ACKNOWLEDGMENT

This work was supported by the U.S. National Science Foundation (NSF) under grant OAC-1911229 and CNS-2009057.

REFERENCES

- [1] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

- [2] Z. Chen, Z. Xu, and D. Wang, "Deep transfer tensor decomposition with orthogonal constraint for recommender systems," 2021.
- [3] T. G. Kolda and D. Hong, "Stochastic gradients for large-scale tensor decomposition," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1066–1095, 2020.
- [4] M. Mondelli and A. Montanari, "On the connection between learning two-layer neural networks and tensor decomposition," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1051–1060.
- [5] Z. Cheng, B. Li, Y. Fan, and Y. Bao, "A novel rank selection scheme in tensor ring decomposition based on reinforcement learning for deep neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3292–3296.
- [6] S. Fernandes, H. Fanaee-T, and J. Gama, "Tensor decomposition for analysing time-evolving social networks: An overview," *Artificial Intelligence Review*, pp. 1–26, 2020.
- [7] Y. Taguchi, "Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data," *BMC bioinformatics*, vol. 19, no. 13, pp. 27–42, 2019.
- [8] F. Wen, H. C. So, and H. Wymeersch, "Tensor decomposition-based beamspace esprit algorithm for multidimensional harmonic retrieval," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4572–4576.
- [9] D. Hong, T. G. Kolda, and J. A. Dueresch, "Generalized canonical polyadic tensor decomposition," *SIAM Review*, vol. 62, no. 1, pp. 133–163, 2020.
- [10] J. Bennett, C. Elkan, B. Liu, P. Smyth, and D. Tikk, "Kdd cup and workshop 2007," *SIGKDD Explor. Newsl.*, vol. 9, no. 2, p. 51–52, Dec. 2007. [Online]. Available: <https://doi.org/10.1145/1345448.1345459>
- [11] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell, "Toward an architecture for never-ending language learning," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, ser. AAAI'10. AAAI Press, 2010, p. 1306–1313.
- [12] S. A. McKee, "Reflections on the memory wall," in *Proceedings of the 1st Conference on Computing Frontiers*, ser. CF '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 162. [Online]. Available: <https://doi.org/10.1145/977091.977115>
- [13] S. Volos, D. Jevdjic, B. Falsafi, and B. Grot, "An effective dram cache architecture for scale-out servers," Tech. Rep., April 2016.
- [14] S. Wijeratne, S. Pattnaik, Z. Chen, R. Kannan, and V. Prasanna, "Programmable fpga-based memory controller," 2021.
- [15] A. S. Gil, J. B. Benitez, M. H. Calviño, and E. H. Gómez, "Reconfigurable cache implemented on an fpga," in *2010 International Conference on Reconfigurable Computing and FPGAs*, 2010, pp. 250–255.
- [16] M. D. Lam, E. E. Rothberg, and M. E. Wolf, "The cache performance and optimizations of blocked algorithms," in *Proceedings of the Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS IV. New York, NY, USA: Association for Computing Machinery, 1991, p. 63–74. [Online]. Available: <https://doi.org/10.1145/106972.106981>
- [17] X. Ma, D. Zhang, and D. Chiou, "Fpga-accelerated transactional execution of graph workloads," in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 227–236. [Online]. Available: <https://doi.org/10.1145/3020078.3021743>
- [18] F. Kjolstad, P. Ahrens, S. Kamil, and S. Amarasinghe, "Tensor algebra compilation with workspaces," in *2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, 2019, pp. 180–192.
- [19] B. W. Bader and T. G. Kolda, "Efficient MATLAB computations with sparse and factored tensors," *SIAM Journal on Scientific Computing*, vol. 30, no. 1, pp. 205–231, December 2007.
- [20] I. Nisa, J. Li, A. Sukumaran-Rajam, R. Vuduc, and P. Sadayappan, "Load-balanced sparse mttkrp on gpus," in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2019, pp. 123–133.
- [21] M. Asiatci and P. Ienne, "Stop crying over your cache miss rate: Handling efficiently thousands of outstanding misses in fpgas," in *Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 310–319. [Online]. Available: <https://doi.org/10.1145/3289602.3293901>
- [22] M. Asiatci and P. Ienne, "Dynaburst: Dynamically assembling dram bursts over a multitude of random accesses," in *2019 29th International Conference on Field Programmable Logic and Applications (FPL)*, 2019, pp. 254–262.
- [23] S. Aananthakrishnan, N. K. Ahmed, V. Cavé, M. Cintra, Y. Demir, K. D. Bois, S. Eyerhan, J. B. Fryman, I. Ganey, W. Heirman, H. Hoppe, J. Howard, I. Hur, M. Kodiyath, S. Jain, D. S. Klowden, M. M. Landowski, L. Montigny, A. More, P. Ossowski, R. Pawlowski, N. Pepperling, F. Petrini, M. Sikora, B. Seshasayee, S. Smith, S. Szkoda, S. Tayal, J. J. Tithi, Y. Vandiessche, and I. P. Wrosz, "PIUMA: programmable integrated unified memory architecture," *CoRR*, vol. abs/2010.06277, 2020. [Online]. Available: <https://arxiv.org/abs/2010.06277>
- [24] N. Srivastava, H. Rong, P. Barua, G. Feng, H. Cao, Z. Zhang, D. Albonese, V. Sarkar, W. Chen, P. Petersen, G. Lowney, A. Herr, C. Hughes, T. Mattson, and P. Dubey, "T2s-tensor: Productively generating high-performance spatial hardware for dense tensor computations," in *2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2019, pp. 181–189.
- [25] N. Srivastava, H. Jin, S. Smith, H. Rong, D. Albonese, and Z. Zhang, "Tensaurus: A versatile accelerator for mixed sparse-dense tensor computations," in *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2020, pp. 689–702.
- [26] R. Zhang, S. Wijeratne, Y. Yang, S. R. Kuppannagari, and V. K. Prasanna, "The hardware implementation of high throughput parallel hash table on fpga using xor-based memory," <https://github.com/pgroupATusc/XOR-hash>.
- [27] —, "A high throughput parallel hash table on fpga using xor-based memory," in *2020 IEEE High Performance Extreme Computing Conference (HPEC)*, 2020, pp. 1–7.
- [28] Xilinx, "Alveo u250 data center accelerator card," <https://www.xilinx.com/products/boards-and-kits/alveo/u250.html>.
- [29] —, "Ultrascale architecture-based fpgas memory ip v1.4," https://www.xilinx.com/support/documentation/ip_documentation/ultrascale/memory_ip/v1_4/pg150-ultrascale-memory-ip.pdf.
- [30] Intel, "External memory interfaces intel arria 10 fpga ip user guide," <https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/ug/ug-20115.pdf>.