Design of a high-performance tensor-matrix multiplication with BLAS

Cem Savaş Başsoy[a,*]

[a] *Hamburg University of Technology, Schwarzenbergstrasse 95, 21071, Hamburg, Germany*

Abstract

The tensor-matrix multiplication is a basic tensor operation required by various tensor methods such as the HOSVD. This paper presents flexible high-performance algorithms that compute the tensor-matrix product according to the Loops-over-GEMM (LoG) approach. Our algorithms are able to process dense tensors with any linear tensor layout, arbitrary tensor order and dimensions all of which can be runtime variable. We discuss two slicing methods with orthogonal parallelization strategies and propose four algorithms that call BLAS with subtensors or tensor slices. We provide a simple heuristic which selects one of the four proposed algorithms at runtime. All algorithms have been evaluated on a large set of tensors with various tensor shapes and linear tensor layouts. In case of large tensor slices, our best-performing algorithm achieves a median performance of 2.47 TFLOPS on an Intel Xeon Gold 5318Y and 2.93 TFLOPS an AMD EPYC 9354. Furthermore, it outperforms batched GEMM implementation of Intel MKL by a factor of 2.57 with large tensor slices. For the majority of our test tensors, our implementation is on average 25.05% faster than other state-of-the-art approaches, including actively developed libraries like Libtorch and Eigen. This work is an extended version of the article "Fast and Layout-Oblivious Tensor-Matrix Multiplication with BLAS" (Bassoy, 2024)[1].

1. Introduction

Tensor computations are found in many scientific fields such as computational neuroscience, pattern recognition, signal processing and data mining [2, 3]. These computations use basic tensor operations as building blocks for decomposing and analyzing multidimensional data which are represented by tensors [4, 5]. Tensor contractions are an important subset of basic operations that need to be fast for efficiently solving tensor methods.

There are three main approaches for implementing tensor contractions. The Transpose Transpose GEMM Transpose (TGGT) approach reorganizes tensors in order to perform a tensor contraction using optimized implementations of the general matrix multiplication (GEMM) [6, 7]. GEMM-like Tensor-Tensor multiplication (GETT) method implement macro-kernels that are similar to the ones used in fast GEMM implementations [8, 9]. The third method is the Loops-over-GEMM (LoG) or the BLAS-based approach in which Basic Linear Algebra Subprograms (BLAS) are utilized with multiple tensor slices or subtensors if possible [10, 11, 12, 13]. The BLAS are considered the de facto standard for writing efficient and portable linear algebra software, which is why nearly all processor vendors provide highly optimized BLAS implementations. Implementations of the LoG and TTGT approaches are in general easier to maintain and faster to port than GETT implementations which might need to adapt vector instructions or blocking parameters according to a processor's microarchitecture.

In this work, we present high-performance algorithms for the tensor-matrix multiplication which is used in many numerical methods such as the alternating least squares method [4, 5]. It is a compute-bound tensor operation and has the same arithmetic intensity as a matrix-matrix multiplication which can almost reach the practical peak performance of a computing machine. To our best knowledge, we are the first to combine the LoG-approach described in [13, 14] for tensor-vector multiplications with the findings on tensor slicing for the tensor-matrix multiplication in [11]. Our algorithms support dense tensors with any order, dimensions and any linear tensor layout including the first- and the last-order storage formats for any contraction mode all of which can be runtime variable. They compute the tensor-matrix product in parallel using efficient GEMM without transposing or flattening tensors. In addition to their high performance, all algorithms are layout-oblivious and provide a sustained performance independent of the tensor layout and without tuning. We provide a single algorithm that selects one of the proposed algorithms based on a simple heuristic.

Every proposed algorithm can be implemented with less than 150 lines of C++ code where the algorithmic complexity is reduced by the BLAS implementation and the corresponding selection of subtensors or tensor slices. We have provided an open-source C++ implementation of all algorithms and a python interface for convenience.

The analysis in this work quantifies the impact of the tensor layout, the tensor slicing method and parallel ex-

*Corresponding author
Email address: `cem.bassoy@gmail.com` (Cem Savaş Başsoy)

ecution of slice-matrix multiplications with varying contraction modes. The runtime measurements of our implementations are compared with state-of-the-art approaches discussed in [8, 9, 15] including Libtorch and Eigen. While our implementation have been benchmarked with the Intel MKL and AMD AOCL libraries, the user choose other BLAS libraries. In summary, the main findings of our work are:

- Given a row-major or column-major input matrix, the tensor-matrix multiplication with tensors of any linear tensor layout can be implemented by an in-place algorithm with 1 GEMV and 7 GEMM instances, supporting all combinations of contraction mode, tensor order and tensor dimensions.
- The proposed algorithms show a similar performance characteristic across different tensor layouts, provided that the contraction conditions remain the same.
- A simple heuristic is sufficient to select one of the proposed algorithms at runtime, providing a near-optimal performance for a wide range of tensor shapes.
- Our best-performing algorithm is a factor of 2.57 faster than Intel's batched GEMM implementation for large tensor slices.
- Our best-performing algorithm is on average 25.05% faster than other state-of-the art library implementations, including LibTorch and Eigen.

The remainder of the paper is organized as follows. Section 2 presents related work. Section 3 introduces some notation on tensors and defines the tensor-matrix multiplication. Algorithm design and methods for slicing and parallel execution are discussed in Section 4. Section 5 describes the test setup. Benchmark results are presented in Section 6. Conclusions are drawn in Section 7.

2. Related Work

Springer et al. [8] present a tensor-contraction generator TCCG and the GETT approach for dense tensor contractions that is inspired from the design of a high-performance GEMM. Their unified code generator selects implementations from generated GETT, LoG and TTGT candidates. Their findings show that among 48 different contractions 15% of LoG-based implementations are the fastest.

Matthews [9] presents a runtime flexible tensor contraction library that uses GETT approach as well. He describes block-scatter-matrix algorithm which uses a special layout for the tensor contraction. The proposed algorithm yields results that feature a similar runtime behavior to those presented in [8].

Li et al. [11] introduce InTensLi, a framework that generates in-place tensor-matrix multiplication according to the LOG approach. The authors discusses optimization and tuning techniques for slicing and parallelizing the operation. With optimized tuning parameters, they report a speedup of up to 4x over the TTGT-based MATLAB tensor toolbox library discussed in [6].

Başsoy [13] presents LoG-based algorithms that compute the tensor-vector product. They support dense tensors with linear tensor layouts, arbitrary dimensions and tensor order. The presented approach is to divide into eight TTV cases calling GEMV and DOT. He reports average speedups of 6.1x and 4.0x compared to implementations that use the TTGT and GETT approach, respectively.

Pawlowski et al. [14] propose morton-ordered blocked layout for a mode-oblivious performance of the tensor-vector multiplication. Their algorithm iterate over blocked tensors and perform tensor-vector multiplications on blocked tensors. They are able to achieve high performance and mode-oblivious computations.

3. Background

3.1. Tensor Notation

An order-p tensor is a p-dimensional array where tensor elements are contiguously stored in memory[16, 4]. We write a, $\mathbf{a}$, $\mathbf{A}$ and $\underline{\mathbf{A}}$ in order to denote scalars, vectors, matrices and tensors. If not otherwise mentioned, we assume $\underline{\mathbf{A}}$ to have order $p > 2$. The p-tuple $\mathbf{n} = (n_1, n_2, \dots, n_p)$ will be referred to as the shape or dimension tuple of a tensor where $n_r > 1$. We will use round brackets $\underline{\mathbf{A}}(i_1, i_2, \dots, i_p)$ or $\underline{\mathbf{A}}(\mathbf{i})$ to denote a tensor element where $\mathbf{i} = (i_1, i_2, \dots, i_p)$ is a multi-index. For convenience, we will also use square brackets to concatenate index tuples such that $[\mathbf{i}, \mathbf{j}] = (i_1, i_2, \dots, i_r, j_1, j_2, \dots, j_q)$ where $\mathbf{i}$ and $\mathbf{j}$ are multi-indices of length r and q, respectively.

3.2. Tensor-Matrix Multiplication

Let $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ be order-p tensors with shapes $\mathbf{n}_a = ([\mathbf{n}_1, n_q, \mathbf{n}_2])$ and $\mathbf{n}_c = ([\mathbf{n}_1, m, \mathbf{n}_2])$ where $\mathbf{n}_1 = (n_1, n_2, \dots, n_{q-1})$ and $\mathbf{n}_2 = (n_{q+1}, n_{q+2}, \dots, n_p)$. Let $\mathbf{B}$ be a matrix of shape $\mathbf{n}_b = (m, n_q)$. A q-mode tensor-matrix product is denoted by $\underline{\mathbf{C}} = \underline{\mathbf{A}} \times_q \mathbf{B}$. An element of $\underline{\mathbf{C}}$ is defined by

$$\underline{\mathbf{C}}([\mathbf{i}_1, j, \mathbf{i}_2]) = \sum_{i_q=1}^{n_q} \underline{\mathbf{A}}([\mathbf{i}_1, i_q, \mathbf{i}_2]) \cdot \mathbf{B}(j, i_q) \tag{1}$$

with $\mathbf{i}_1 = (i_1, \dots, i_{q-1})$, $\mathbf{i}_2 = (i_{q+1}, \dots, i_p)$ where $1 \le i_r \le n_r$ and $1 \le j \le m$ [11, 5]. The mode q is called the contraction mode with $1 \le q \le p$. The tensor-matrix multiplication generalizes the computational aspect of the two-dimensional case $\mathbf{C} = \mathbf{B} \cdot \mathbf{A}$ if $p = 2$ and $q = 1$. Its arithmetic intensity is equal to that of a matrix-matrix multiplication and is not memory-bound.

In the following, we assume that the tensors $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ have the same tensor layout $\boldsymbol{\pi}$. Elements of matrix $\underline{\mathbf{B}}$ can be stored either in the column-major or row-major format. The tensor-matrix multiplication with i_q iterating over the second mode of $\mathbf{B}$ is also referred to as the q-mode product which is a building block for tensor methods such as the higher-order orthogonal iteration or the higher-order singular value decomposition [5]. Please note that the following method can be applied, if indices j and i_q of matrix $\mathbf{B}$ are swapped.

3.3. Subtensors

A subtensor references elements of a tensor $\underline{\mathbf{A}}$ and is denoted by $\underline{\mathbf{A}}'$. It is specified by a selection grid that consists of p index ranges. In this work, an index range of a given mode r shall either contain all indices of the mode r or a single index i_r of that mode where $1 \leq r \leq p$. Subtensor dimensions n'_r are either n_r if the full index range or 1 if a a single index for mode r is used. Subtensors are annotated by their non-unit modes such as $\underline{\mathbf{A}}'_{u,v,w}$ where $n_u > 1$, $n_v > 1$ and $n_w > 1$ for $1 \leq u \neq v \neq w \leq p$. The remaining single indices of a selection grid can be inferred by the loop induction variables of an algorithm. The number of non-unit modes determine the order p' of subtensor where $1 \leq p' < p$. In the above example, the subtensor $\underline{\mathbf{A}}'_{u,v,w}$ has three non-unit modes and is thus of order 3. For convenience, we might also use an dimension tuple $\mathbf{m}$ of length p' with $\mathbf{m} = (m_1, m_2, \ldots, m_{p'})$ to specify a mode-p' subtensor $\underline{\mathbf{A}}'_{\mathbf{m}}$. An order-2 subtensor of $\underline{\mathbf{A}}'$ is a tensor slice $\mathbf{A}'_{u,v}$ and an order-1 subtensor of $\underline{\mathbf{A}}'$ is a fiber $\mathbf{a}'_u$.

3.4. Linear Tensor Layouts

We use a layout tuple $\boldsymbol{\pi} \in \mathbb{N}^p$ to encode all linear tensor layouts including the first-order or last-order layout. They contain permuted tensor modes whose priority is given by their index. For instance, the general k-order tensor layout for an order-p tensor is given by the layout tuple $\boldsymbol{\pi}$ with $\pi_r = k - r + 1$ for $1 < r \leq k$ and r for $k < r \leq p$. The first- and last-order storage formats are given by $\boldsymbol{\pi}_F = (1, 2, \ldots, p)$ and $\boldsymbol{\pi}_L = (p, p-1, \ldots, 1)$. An inverse layout tuple $\boldsymbol{\pi}^{-1}$ is defined by $\boldsymbol{\pi}^{-1}(\boldsymbol{\pi}(k)) = k$. Given a layout tuple $\boldsymbol{\pi}$ with p modes, the π_r-th element of a stride tuple is given by $w_{\pi_r} = \prod_{k=1}^{r-1} n_{\pi_k}$ for $1 < r \leq p$ and $w_{\pi_1} = 1$. Tensor elements of the π_1-th mode are contiguously stored in memory. The location of tensor elements is determined by the tensor layout and the layout function. For a given tensor layout and stride tuple, a layout function $\lambda_{\mathbf{w}}$ maps a multi-index to a scalar index with $\lambda_{\mathbf{w}}(\mathbf{i}) = \sum_{r=1}^{p} w_r(i_r - 1)$, see [17, 14].

3.5. Flattening and Reshaping

The following two operations define non-modifying reformatting transformations of dense tensors with contiguously stored elements and linear tensor layouts.

The flattening operation $\varphi_{u,v}$ transforms an order-p tensor $\underline{\mathbf{A}}$ with a shape $\mathbf{n}$ and layout $\boldsymbol{\pi}$ tuple to an order-p' view $\underline{\mathbf{B}}$ with a shape $\mathbf{m}$ and layout $\boldsymbol{\tau}$ tuple of length p' with $p' = p - v + u$ and $1 \leq u < v \leq p$. It is akin to tensor unfolding, also known as matricization and vectorization [5, p.459]. However, it neither modifies the element ordering nor copies tensor elements. Given a layout tuple $\boldsymbol{\pi}$ of $\underline{\mathbf{A}}$, the flattening operation $\varphi_{u,v}$ is defined for contiguous modes $\hat{\boldsymbol{\pi}} = (\pi_u, \pi_{u+1}, \ldots, \pi_v)$ of $\boldsymbol{\pi}$. With $j_k = 0$ if $k \leq u$ and $j_k = v - u$ if $k > u$ where $1 \leq k \leq p'$, the resulting layout tuple $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{p'})$ of $\underline{\mathbf{B}}$ is then given by $\tau_u = \min(\boldsymbol{\pi}_{u,v})$ and $\tau_k = \pi_{k+j_k} - s_k$ for $k \neq u$ with $s_k = |\{\pi_i \mid \pi_{k+j_k} > \pi_i \wedge \pi_i \neq \min(\hat{\boldsymbol{\pi}}) \wedge u \leq i \leq p\}|$. Elements of the shape tuple $\mathbf{m}$ are defined by $m_{\tau_u} = \prod_{k=u}^{v} n_{\pi_k}$ and $m_{\tau_k} = n_{\pi_{k+j}}$ for $k \neq u$.

4. Algorithm Design

4.1. Baseline Algorithm with Contiguous Memory Access

The tensor-times-matrix multiplication in equation 1 can be implemented with one sequential algorithm using a nested recursion [17]. It consists of two `if` statements with an `else` branch that computes a fiber-matrix product with two loops. The outer loop iterates over the dimension m of $\underline{\mathbf{C}}$ and $\mathbf{B}$, while the inner iterates over dimension n_q of $\underline{\mathbf{A}}$ and $\mathbf{B}$ computing an inner product with fibers of $\underline{\mathbf{A}}$ and $\mathbf{B}$. While matrix $\mathbf{B}$ can be accessed contiguously depending on its storage format, elements of $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ are accessed non-contiguously if $\pi_1 \neq q$.

A better approach is illustrated in algorithm 1 where the loop order is adjusted to the tensor layout $\boldsymbol{\pi}$ and memory is accessed contiguously for $\pi_1 \neq q$ and $p > 1$. The rearrangement of the loop order is accomplished in line 5 which uses the layout tuple $\boldsymbol{\pi}$ to select a multi-index element i_{π_r} and to increment it with the corresponding stride w_{π_r}. Hence, with increasing recursion level and decreasing r, indices are incremented with smaller strides as $w_{\pi_r} \leq w_{\pi_{r+1}}$. The second `if` statement in line number 4 allows the loop over mode π_1 to be placed into the base case which contains three loops performing a slice-matrix multiplication. In this way, the inner-most loop is able to increment i_{π_1} with a unit stride and contiguously accesses tensor elements of $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$. The second loop increments i_q with which elements of $\mathbf{B}$ are contiguously accessed if $\mathbf{B}$ is stored in the row-major format. The third loop increments j and could be placed as the second loop if $\mathbf{B}$ is stored in the column-major format.

While spatial data locality is improved by adjusting the loop ordering, slices $\underline{\mathbf{A}}'_{\pi_1,q}$, fibers $\underline{\mathbf{C}}'_{\pi_1}$ and elements $\underline{\mathbf{B}}(j, i_q)$ are accessed m, n_q and n_{π_1} times, respectively. The specified fiber of $\underline{\mathbf{C}}$ might fit into first or second level cache, slice elements of $\underline{\mathbf{A}}$ are unlikely to fit in the local caches if the slice size $n_{\pi_1} \times n_q$ is large, leading to higher cache misses and suboptimal performance. Instead of attempting to improve the temporal data locality, we make use of existing high-performance BLAS implementations for the base case. The following subsection explains this approach.

$\texttt{ttm}(\underline{\mathbf{A}}, \mathbf{B}, \underline{\mathbf{C}}, \mathbf{n}, \boldsymbol{\pi}, \mathbf{i}, m, q, \hat{q}, r)$
  if $r = \hat{q}$ **then**
    $\texttt{ttm}(\underline{\mathbf{A}}, \mathbf{B}, \underline{\mathbf{C}}, \mathbf{n}, \boldsymbol{\pi}, \mathbf{i}, m, q, \hat{q}, r-1)$
  else if $r > 1$ **then**
    for $i_{\pi_r} \leftarrow 1$ **to** n_{π_r} **do**
      $\texttt{ttm}(\underline{\mathbf{A}}, \mathbf{B}, \underline{\mathbf{C}}, \mathbf{n}, \boldsymbol{\pi}, \mathbf{i}, m, q, \hat{q}, r-1)$
  else
    for $j \leftarrow 1$ **to** m **do**
      for $i_q \leftarrow 1$ **to** n_q **do**
        for $i_{\pi_1} \leftarrow 1$ **to** n_{π_1} **do**
          $\underline{\mathbf{C}}([\mathbf{i}_1, j, \mathbf{i}_2])$ += $\underline{\mathbf{A}}([\mathbf{i}_1, i_q, \mathbf{i}_2]) \cdot \mathbf{B}(j, i_q)$

Algorithm 1: Modified baseline algorithm with contiguous memory access for the tensor-matrix multiplication. The tensor order p must be greater than 1 and the contraction mode q must satisfy $1 \leq q \leq p$ and $\pi_1 \neq q$. The initial call must happen with $r = p$ where $\mathbf{n}$ is the shape tuple of $\underline{\mathbf{A}}$ and m is the q-th dimension of $\underline{\mathbf{C}}$.

4.2. BLAS-based Algorithms with Tensor Slices

The following approach utilizes the CBLAS `gemm` function in the base case of Algorithm 1 in order to perform fast slice-matrix multiplications[1]. Function `gemm` denotes a general matrix-matrix multiplication which is defined as `C:=a*op(A)*op(B)+b*C` where `a` and `b` are scalars, `A`, `B` and `C` are matrices, `op(A)` is an `M`-by-`K` matrix, `op(B)` is a `K`-by-`N` matrix and `C` is an `N`-by-`N` matrix. Function `op(x)` either transposes the corresponding matrix `x` such that `op(x)=x'` or not `op(x)=x`. The CBLAS interface also allows users to specify matrix's leading dimension by providing the `LDA`, `LDB` and `LDC` parameters. A leading dimension specifies the number of elements that is required for iterating over the non-contiguous matrix dimension. The leading dimension can be used to perform a matrix multiplication with submatrices or even fibers within submatrices. The leading dimension parameter is necessary for the BLAS-based tensor-matrix multiplication.

The eighth TTM case in Table 1 contains all arguments that are necessary to perform a CBLAS `gemm` in the base case of Algorithm 1. The arguments of `gemm` are set according to the tensor order p, tensor layout $\boldsymbol{\pi}$ and contraction mode q. If the input matrix $\mathbf{B}$ has the row-major order, parameter `CBLAS_ORDER` of function `gemm` is set to `CblasRowMajor` (`rm`) and `CblasColMajor` (`cm`) otherwise. The eighth case will be denoted as the general case in which function `gemm` is called multiple times with different tensor slices. Next to the eighth TTM case, there are seven corner cases where a single `gemv` or `gemm` call suffices to compute the tensor-matrix product. For instance if $\pi_1 = q$, the tensor-matrix product can be computed by a matrix-matrix multiplication where the input tensor $\underline{\mathbf{A}}$ can be flattened into a matrix without any copy operation. Note that Table 1 supports all linear tensor layouts of $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ with no limitations on tensor order and contraction mode. The following subsection describes all eight TTM cases when the input matrix $\mathbf{B}$ has the row-major ordering.

4.2.1. Row-Major Matrix Multiplication

The following paragraphs introduce all TTM cases that are listed in Table 1.

Case 1: If $p = 1$, The tensor-vector product $\underline{\mathbf{A}} \times_1 \mathbf{B}$ can be computed with a `gemv` operation where $\underline{\mathbf{A}}$ is an order-1 tensor $\mathbf{a}$ of length n_1 such that $\mathbf{a}^T \cdot \mathbf{B}$.

Case 2-5: If $p = 2$, $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ are order-2 tensors with dimensions n_1 and n_2. In this case the tensor-matrix product can be computed with a single `gemm`. If $\mathbf{A}$ and $\mathbf{C}$ have the column-major format with $\boldsymbol{\pi} = (1, 2)$, `gemm` either executes $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}^T$ for $q = 1$ or $\mathbf{C} = \mathbf{B} \cdot \mathbf{A}$ for $q = 2$. Both matrices can be interpreted $\mathbf{C}$ and $\mathbf{A}$ as matrices in row-major format although both are stored column-wise. If $\mathbf{A}$ and $\mathbf{C}$ have the row-major format with $\boldsymbol{\pi} = (2, 1)$, `gemm` either executes $\mathbf{C} = \mathbf{B} \cdot \mathbf{A}$ for $q = 1$ or $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}^T$ for $q = 2$. The transposition of $\mathbf{B}$ is necessary for the TTM cases 2 and 5 which is independent of the chosen layout.

Case 6-7 : If $p > 2$ and if $q = \pi_1$(case 6), a single `gemm` with the corresponding arguments executes $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}^T$ and computes a tensor-matrix product $\underline{\mathbf{C}} = \underline{\mathbf{A}} \times_{\pi_1} \mathbf{B}$. Tensors $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ are flattened with $\varphi_{2,p}$ to row-major matrices $\mathbf{A}$ and $\mathbf{C}$. Matrix $\mathbf{A}$ has $\bar{n}_{\pi_1} = \bar{n}/n_{\pi_1}$ rows and n_{π_1} columns while matrix $\mathbf{C}$ has the same number of rows and m columns. If $\pi_p = q$ (case 7), $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ are flattened with $\varphi_{1,p-1}$ to column-major matrices $\mathbf{A}$ and $\mathbf{C}$. Matrix $\mathbf{A}$ has n_{π_p} rows and $\bar{n}_{\pi_p} = \bar{n}/n_{\pi_p}$ columns while $\mathbf{C}$ has m rows and the same number of columns. In this case, a single `gemm` executes $\mathbf{C} = \mathbf{B} \cdot \mathbf{A}$ and computes $\underline{\mathbf{C}} = \underline{\mathbf{A}} \times_{\pi_p} \mathbf{B}$. Noticeably, the desired contraction are performed without copy operations, see subsection 3.5.

Case 8 ($p > 2$): If the tensor order is greater than 2 with $\pi_1 \neq q$ and $\pi_p \neq q$, the modified baseline algorithm 1 is used to successively call $\bar{n}/(n_q \cdot n_{\pi_1})$ times `gemm` with different tensor slices of $\underline{\mathbf{C}}$ and $\underline{\mathbf{A}}$. Each `gemm` computes one slice $\underline{\mathbf{C}}'_{\pi_1,q}$ of the tensor-matrix product $\underline{\mathbf{C}}$ using the corresponding tensor slices $\underline{\mathbf{A}}'_{\pi_1,q}$ and the matrix $\mathbf{B}$. The matrix-matrix product $\mathbf{C} = \mathbf{B} \cdot \mathbf{A}$ is performed by interpreting both tensor slices as row-major matrices $\mathbf{A}$ and $\mathbf{C}$ which have the dimensions (n_q, n_{π_1}) and (m, n_{π_1}), respectively.

4.2.2. Column-Major Matrix Multiplication

The tensor-matrix multiplication is performed with the column-major version of `gemm` when the input matrix $\mathbf{B}$ is stored in column-major order. Although the number of `gemm` cases remains the same, the `gemm` arguments must be rearranged. The argument arrangement for the column-major version can be derived from the row-major version that is provided in table 1.

The CBLAS arguments of `M` and `N`, as well as `A` and `B` is swapped and the transposition flag for matrix $\mathbf{B}$ is toggled. Also, the leading dimension argument of `A` is adjusted to `LDB` or `LDA`. The only new argument is the new leading dimension of `B`.

[1] CBLAS denotes the C interface to the BLAS.

Case	Order p	Layout $\boldsymbol{\pi}_{\underline{\mathbf{A}},\underline{\mathbf{C}}}$	Layout $\boldsymbol{\pi}_{\mathbf{B}}$	Mode q	Routine	`T`	`M`	`N`	`K`	`A`	`LDA`	`B`	`LDB`	`LDC`
1	1	-	`rm/cm`	1	`gemv`	-	m	n_1	-	$\mathbf{B}$	n_1	$\underline{\mathbf{A}}$	-	-
2	2	`cm`	`rm`	1	`gemm`	$\mathbf{B}$	n_2	m	n_1	$\underline{\mathbf{A}}$	n_1	$\mathbf{B}$	n_1	m
	2	`cm`	`cm`	1	`gemm`	-	m	n_2	n_1	$\mathbf{B}$	m	$\underline{\mathbf{A}}$	n_1	m
3	2	`cm`	`rm`	2	`gemm`	-	m	n_1	n_2	$\mathbf{B}$	n_2	$\underline{\mathbf{A}}$	n_1	n_1
	2	`cm`	`cm`	2	`gemm`	$\mathbf{B}$	n_1	m	n_2	$\underline{\mathbf{A}}$	n_1	$\mathbf{B}$	m	n_1
4	2	`rm`	`rm`	1	`gemm`	-	m	n_2	n_1	$\mathbf{B}$	n_1	$\underline{\mathbf{A}}$	n_2	n_2
	2	`rm`	`cm`	1	`gemm`	$\mathbf{B}$	n_2	m	n_1	$\underline{\mathbf{A}}$	n_2	$\mathbf{B}$	m	n_2
5	2	`rm`	`rm`	2	`gemm`	$\mathbf{B}$	n_1	m	n_2	$\underline{\mathbf{A}}$	n_2	$\mathbf{B}$	n_2	m
	2	`rm`	`cm`	2	`gemm`	-	m	n_1	n_2	$\mathbf{B}$	m	$\underline{\mathbf{A}}$	n_2	m
6	> 2	any	`rm`	π_1	`gemm`	$\mathbf{B}$	$\bar{n}_q$	m	n_q	$\underline{\mathbf{A}}$	n_q	$\mathbf{B}$	n_q	m
	> 2	any	`cm`	π_1	`gemm`	-	m	$\bar{n}_q$	n_q	$\mathbf{B}$	m	$\underline{\mathbf{A}}$	n_q	m
7	> 2	any	`rm`	π_p	`gemm`	-	m	$\bar{n}_q$	n_q	$\mathbf{B}$	n_q	$\underline{\mathbf{A}}$	$\bar{n}_q$	$\bar{n}_q$
	> 2	any	`cm`	π_p	`gemm`	$\mathbf{B}$	$\bar{n}_q$	m	n_q	$\underline{\mathbf{A}}$	$\bar{n}_q$	$\mathbf{B}$	m	$\bar{n}_q$
8	> 2	any	`rm`	$\pi_2, .., \pi_{p-1}$	`gemm*`	-	m	n_{π_1}	n_q	$\mathbf{B}$	n_q	$\underline{\mathbf{A}}$	w_q	w_q
	> 2	any	`cm`	$\pi_2, .., \pi_{p-1}$	`gemm*`	$\mathbf{B}$	n_{π_1}	m	n_q	$\underline{\mathbf{A}}$	w_q	$\mathbf{B}$	m	w_q

Table 1: Eight TTM cases implementing the mode-q tensor-matrix multiplication with the `gemm` and `gemv` CBLAS functions. Arguments of `gemv` and `gemm` (`T`, `M`, `N`, *dots*) are chosen with respect to the tensor order p, layout $\boldsymbol{\pi}$ of $\underline{\mathbf{A}}$, $\mathbf{B}$, $\underline{\mathbf{C}}$ and contraction mode q where `T` specifies if $\mathbf{B}$ is transposed. Function `gemm*` with a star denotes multiple `gemm` calls with different tensor slices. Argument $\bar{n}_q$ for case 6 and 7 is defined as $\bar{n}_q = (\prod_r^p n_r)/n_q$. Input matrix $\mathbf{B}$ is either stored in the column-major or row-major format. The storage format flag set for `gemm` and `gemv` is determined by the element ordering of $\mathbf{B}$.

Given case 4 with the row-major matrix multiplication in Table 1 where tensor $\underline{\mathbf{A}}$ and matrix $\mathbf{B}$ are passed to `B` and `A`. The corresponding column-major version is attained when tensor $\underline{\mathbf{A}}$ and matrix $\mathbf{B}$ are passed to `A` and `B` where the transpose flag for $\mathbf{B}$ is set and the remaining dimensions are adjusted accordingly.

4.2.3. Matrix Multiplication Variations

The column-major and row-major versions of `gemm` can be used interchangeably by adapting the storage format. This means that a `gemm` operation for column-major matrices can compute the same matrix product as one for row-major matrices, provided that the arguments are rearranged accordingly. While the argument rearrangement is similar, the arguments associated with the matrices `A` and `B` must be interchanged. Specifically, `LDA` and `LDB` as well as `M` and `N` are swapped along with the corresponding matrix pointers. In addition, the transposition flag must be set for `A` or `B` in the new format if `B` or `A` is transposed in the original version.

For instance, the column-major matrix multiplication in case 4 of table 1 requires the arguments of `A` and `B` to be tensor $\underline{\mathbf{A}}$ and matrix $\mathbf{B}$ with $\mathbf{B}$ being transposed. The arguments of an equivalent row-major multiplication for `A`, `B`, `M`, `N`, `LDA`, `LDB` and `T` are then initialized with $\mathbf{B}$, $\underline{\mathbf{A}}$, m, n_2, m, n_2 and $\mathbf{B}$.

Another possible matrix multiplication variant with the same product is computed when, instead of $\mathbf{B}$, tensors $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ with adjusted arguments are transposed. We assume that such reformulations of the matrix multiplication do not outperform the variants shown in Table 1, as we expect BLAS libraries to have optimal blocking and multiplication strategies.

4.3. Matrix Multiplication with Subtensors

Algorithm 1 can be slightly modified in order to call `gemm` with flattened order-$\hat{q}$ subtensors that correspond to larger tensor slices. Given the contraction mode q with $1 < q < p$, the maximum number of additionally fusible modes is $\hat{q} - 1$ with $\hat{q} = \boldsymbol{\pi}^{-1}(q)$ where $\boldsymbol{\pi}^{-1}$ is the inverse layout tuple. The corresponding fusible modes are therefore $\pi_1, \pi_2, \ldots, \pi_{\hat{q}-1}$.

The non-base case of the modified algorithm only iterates over dimensions that have indices larger than $\hat{q}$ and thus omitting the first $\hat{q}$ modes. The conditions in line 2 and 4 are changed to $1 < r \leq \hat{q}$ and $\hat{q} < r$, respectively. Thus, loop indices belonging to the outer π_r-th loop with $\hat{q}+1 \leq r \leq p$ define the order-$\hat{q}$ subtensors $\underline{\mathbf{A}}'_{\boldsymbol{\pi}'}$ and $\underline{\mathbf{C}}'_{\boldsymbol{\pi}'}$ of $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ with $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_{\hat{q}-1}, q)$. Flattening the subtensors $\underline{\mathbf{A}}'_{\boldsymbol{\pi}'}$ and $\underline{\mathbf{C}}'_{\boldsymbol{\pi}'}$ with $\varphi_{1,\hat{q}-1}$ for the modes $\pi_1, \ldots, \pi_{\hat{q}-1}$ yields two tensor slices with dimension n_q or m with the fused dimension $\bar{n}_q = \prod_{r=1}^{\hat{q}-1} n_{\pi_r}$ and $\bar{n}_q = w_q$. Both tensor slices can be interpreted either as row-major or column-major matrices with shapes $(n_q, \bar{n}_q)$ or $(w_q, \bar{n}_q)$ in case of $\underline{\mathbf{A}}$ and $(m, \bar{n}_q)$ or $(\bar{n}_q, m)$ in case of $\underline{\mathbf{C}}$, respectively.

The `gemm` function in the base case is called with almost identical arguments except for the parameter M or N which is set to $\bar{n}_q$ for a column-major or row-major multiplication, respectively. Note that neither the selection of the subtensor nor the flattening operation copy tensor elements. This description supports all linear tensor layouts and generalizes lemma 4.2 in [11] without copying tensor elements, see section 3.5.

4.4. Parallel BLAS-based Algorithms

Most BLAS libraries provide an option to change the number of threads. Hence, functions such as `gemm` and `gemv` can be run either using a single or multiple threads. The

ttm<par-loop><slice>($\underline{\mathbf{A}}$, $\mathbf{B}$, $\underline{\mathbf{C}}$, $\mathbf{n}$, $\boldsymbol{\pi}$, m, q, p)
    [$\underline{\mathbf{A}}'$, $\underline{\mathbf{C}}'$, $\mathbf{n}'$, $\mathbf{w}'$] = flatten ($\underline{\mathbf{A}}$, $\underline{\mathbf{C}}$, $\mathbf{n}$, m, $\boldsymbol{\pi}$, q, p)
    parallel for $i \leftarrow 1$ **to** n'_4 **do**
        parallel for $j \leftarrow 1$ **to** n'_2 **do**
            gemm(m, n'_1, n'_3, 1, $\mathbf{B}$,n'_3, $\underline{\mathbf{A}}'_{ij}$,$w'_3$, 0, $\underline{\mathbf{C}}'_{ij}$,$w'_3$)

Algorithm 2: Function ttm<par-loop><slice> is an optimized version of Algorithm 1. The flatten function transforms the order-p tensors $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ with layout tuple $\boldsymbol{\pi}$ and their respective dimension tuples $\mathbf{n}$ and $\mathbf{m}$ into order-4 tensors $\underline{\mathbf{A}}'$ and $\underline{\mathbf{C}}'$ with layout tuple $\boldsymbol{\pi}'$ and their respective dimension tuples $\mathbf{n}'$ and $\mathbf{m}'$ where $\mathbf{n}' = (n_{\pi_1}, \hat{n}_{\pi_2}, n_q, \hat{n}_{\pi_4})$ and $m'_3 = m$ and $n'_k = m'_k$ for $k \neq 3$. Each thread calls multiple single-threaded gemm functions each of which executes a slice-matrix multiplication with the order-2 tensor slices $\underline{\mathbf{A}}'_{ij}$ and $\underline{\mathbf{C}}'_{ij}$. Matrix $\mathbf{B}$ has the row-major storage format.

TTM cases one to seven contain a single BLAS call which is why we set the number of threads to the number of available cores. The following subsections discuss parallel versions for the eighth case in which the outer loops of algorithm 1 and the gemm function inside the base case can be run in parallel. Note that the parallelization strategies can be combined with the aforementioned slicing methods.

4.4.1. Sequential Loops and Parallel Matrix Multiplication

Algorithm 1 is run for the eighth case and does not need to be modified except for enabling gemm to run multi-threaded in the base case. This type of parallelization strategy might be beneficial with order-$\hat{q}$ subtensors where the contraction mode satisfies $q = \pi_{p-1}$, the inner dimensions $n_{\pi_1}, \ldots, n_{\hat{q}}$ are large and the outer-most dimension n_{π_p} is smaller than the available processor cores. For instance, given a first-order storage format and the contraction mode q with $q = p - 1$ and $n_p = 2$, the dimensions of flattened order-q subtensors are $\prod_{r=1}^{p-2} n_r$ and n_{p-1}. This allows gemm to perform with large dimensions using multiple threads increasing the likelihood to reach a high throughput. However, if the above conditions are not met, a multi-threaded gemm operates on small tensor slices which might lead to an suboptimal utilization of the available cores. This algorithm version will be referred to as <par-gemm>. Depending on the subtensor shape, we will either add <slice> for order-2 subtensors or <subtensor> for order-$\hat{q}$ subtensors with $\hat{q} = \pi_q^{-1}$.

4.4.2. Parallel Loops and Sequential Matrix Multiplication

Instead of sequentially calling multi-threaded gemm, it is also possible to call single-threaded gemms in parallel. Similar to the previous approach, the matrix multiplication can be performed with tensor slices or order-$\hat{q}$ subtensors.

Matrix Multiplication with Tensor Slices. Algorithm 2 with function ttm<par-loop><slice> executes a single-threaded gemm with tensor slices in parallel using all modes except π_1 and $\pi_{\hat{q}}$. The first statement of the algorithm calls the flatten function which transforms tensors $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ without copying elements by calling the flattening operation $\varphi_{\pi_{\hat{q}+1},\pi_p}$ and $\varphi_{\pi_2,\pi_{\hat{q}-1}}$. The resulting tensors $\underline{\mathbf{A}}'$ and $\underline{\mathbf{C}}'$ are of order 4. Tensor $\underline{\mathbf{A}}'$ has the shape $\mathbf{n}' = (n_{\pi_1}, \hat{n}_{\pi_2}, n_q, \hat{n}_{\pi_4})$ with the dimensions $\hat{n}_{\pi_2} = \prod_{r=2}^{\hat{q}-1} n_{\pi_r}$ and $\hat{n}_{\pi_4} = \prod_{r=\hat{q}+1}^{p} n_{\pi_r}$. Tensor $\underline{\mathbf{C}}'$ has the same shape as $\underline{\mathbf{A}}'$ with dimensions $m'_r = n'_r$ except for the third dimension which is given by $m_3 = m$.

The following two parallel for loops iterate over all free modes. The outer loop iterates over $n'_4 = \hat{n}_{\pi_4}$ while the inner one loops over $n'_2 = \hat{n}_{\pi_2}$ calling gemm with tensor slices $\underline{\mathbf{A}}'_{2,4}$ and $\underline{\mathbf{C}}'_{2,4}$. Here, we assume that matrix $\mathbf{B}$ has the row-major format which is why both tensor slices are also treated as row-major matrices. Notice that gemm in Algorithm 2 will be called with exact same arguments as displayed in the eighth case in Table 1 where $n'_1 = n_{\pi_1}$, $n'_3 = n_q$ and $w_q = w'_3$. For the sake of simplicity, we omitted the first three arguments of gemm which are set to CblasRowMajor and CblasNoTrans for A and B. With the help of the flattening operation, the tree-recursion has been transformed into two loops which iterate over all free indices.

Matrix Multiplication with Subtensors. An alternative algorithm is given by combining Algorithm 2 with order-$\hat{q}$ subtensors that have been discussed in 4.3. With order-$\hat{q}$ subtensors, only the outer modes $\pi_{\hat{q}+1}, \ldots, \pi_p$ are free for parallel execution while the inner modes $\pi_1, \ldots, \pi_{\hat{q}-1}, q$ are used for the slice-matrix multiplication. Therefore, both tensors are flattened twice using the flattening operations $\varphi_{\pi_1,\pi_{\hat{q}-1}}$ and $\varphi_{\pi_{\hat{q}+1},\pi_p}$. Note that in contrast to tensor slices, the first flattening also contains the dimension n_{π_1}. The flattened tensors are of order 3 where $\underline{\mathbf{A}}'$ has the shape $\mathbf{n}' = (\hat{n}_{\pi_1}, n_q, \hat{n}_{\pi_3})$ with $\hat{n}_{\pi_1} = \prod_{r=1}^{\hat{q}-1} n_{\pi_r}$ and $\hat{n}_{\pi_3} = \prod_{r=\hat{q}+1}^{p} n_{\pi_r}$. Tensor $\underline{\mathbf{C}}'$ has the same dimensions as $\underline{\mathbf{A}}'$ except for $m_2 = m$.

Algorithm 2 needs a minor modification for supporting order-$\hat{q}$ subtensors. Instead of two loops, the modified algorithm consists of a single loop which iterates over dimension $\hat{n}_{\pi_3}$ calling a single-threaded gemm with subtensors $\underline{\mathbf{A}}'$ and $\underline{\mathbf{C}}'$. The shape and strides of both subtensors as well as the function arguments of gemm have already been provided by the previous subsection 4.3. This ttm version will referred to as <par-loop><subtensor>.

Note that functions <par-gemm> and <par-loop> implement opposing versions of the ttm where either gemm or the fused loop is performed in parallel. Version <par-loop-gemm executes available loops in parallel where each loop thread executes a multi-threaded gemm with either subtensors or tensor slices.

4.4.3. Combined Matrix Multiplication

The combined matrix multiplication calls one of the previously discussed functions depending on the number of available cores. The heuristic assumes that function <par-gemm> is not able to efficiently utilize the processor cores if subtensors or tensor slices are too small. The corresponding algorithm switches between <par-loop> and <par-gemm> with subtensors by first calculating the parallel and combined loop count $\hat{n} = \prod_{r=1}^{\hat{q}-1} n_{\pi_r}$ and $\hat{n}' = \prod_{r=1}^{p} n_{\pi_r}/n_q$, respectively. Given the number of physical processor cores as ncores, the algorithm executes <par-loop> with <subtensor> if ncores is greater than or equal to $\hat{n}$ and call <par-loop> with <slice> if ncores is greater than or equal to $\hat{n}'$. Otherwise, the algorithm will default to <par-gemm> with <subtensor>. Function par-gemm with tensor slices is not used here. The presented strategy is different to the one presented in [11] that maximizes the number of modes involved in the matrix multiply. We will refer to this version as <combined> to denote a selected combination of <par-loop> and <par-gemm> functions.

4.4.4. Multithreaded Batched Matrix Multiplication

The multithreaded batched matrix multiplication version calls in the eighth case a single gemm_batch function that is provided by Intel MKL's BLAS-like extension. With an interface that is similar to the one of cblas_gemm, function gemm_batch performs a series of matrix-matrix operations with general matrices. All parameters except CBLAS_LAYOUT requires an array as an argument which is why different subtensors of the same corresponding tensors are passed to gemm_batch. The subtensor dimensions and remaining gemm arguments are replicated within the corresponding arrays. Note that the MKL is responsible of how subtensor-matrix multiplications are executed and whether subtensors are further divided into smaller subtensors or tensor slices. This algorithm will be referred to as <mkl-batch-gemm>.

5. Experimental Setup

5.1. Computing System

The experiments have been carried out on a dual socket Intel Xeon Gold 5318Y CPU with an Ice Lake architecture and a dual socket AMD EPYC 9354 CPU with a Zen4 architecture. With two NUMA domains, the Intel CPU consists of 2×24 cores which run at a base frequency of 2.1 GHz. Assuming a peak AVX-512 Turbo frequency of 2.5 GHz, the CPU is able to process 3.84 TFLOPS in double precision. We measured a peak double-precision floating-point performance of 3.8043 TFLOPS (79.25 GFLOPS/core) and a peak memory throughput of 288.68 GB/s using the Likwid performance tool. The AMD EPYC 9354 CPU consists of 2×32 cores running at a base frequency of 3.25 GHz. Assuming an all-core boost frequency of 3.75 GHz, the CPU is theoretically capable of performing 3.84 TFLOPS in double precision. We measured a peak double-precision floating-point performance of 3.87 TFLOPS (60.5 GFLOPS/core) and a peak memory throughput of 788.71 GB/s.

We have used the GNU compiler v11.2.0 with the highest optimization level -O3 together with the -fopenmp and -std=c++17 flags. Loops within the eighth case have been parallelized using GCC's OpenMP v4.5 implementation. In case of the Intel CPU, the 2022 Intel Math Kernel Library (MKL) and its threading library mkl_intel_thread together with the threading runtime library libiomp5 has been used for the three BLAS functions gemv, gemm and gemm_batch. For the AMD CPU, we have compiled AMD AOCL v4.2.0 together with set the zen4 architecture configuration option and enabled OpenMP threading.

5.2. OpenMP Parallelization

The loops in the par-loop algorithms have been parallelized using the OpenMP directive omp parallel for together with the schedule(static), num_threads(ncores) and proc_bind(spread) clauses. In case of tensor-slices, the collapse(2) clause has been added for transforming both loops into one loop which has an iteration space of the first loop times the second one. We also had to enable nested parallelism using omp_set_nested to toggle between single- and multi-threaded gemm calls for different TTM cases when using AMD AOCL.

The num_threads(ncores) clause specifies the number of threads within a team where ncores is equal to the number of processor cores. Hence, each OpenMP thread is responsible for computing $\bar{n}'/$ncores independent slice-matrix products where $\bar{n}' = n_2' \cdot n_4'$ for tensor slices and $\bar{n}' = n_4'$ for mode-$\hat{q}$ subtensors.

The schedule(static) instructs the OpenMP runtime to divide the iteration space into almost equally sized chunks. Each thread sequentially computes $\bar{n}'/$ncores slice-matrix products. We have decided to use this scheduling kind as all slice-matrix multiplications exhibit the same number of floating-point operations with a regular workload where one can assume negligible load imbalance. Moreover, we wanted to prevent scheduling overheads for small slice-matrix products were data locality can be an important factor for achieving higher throughput.

The OMP_PLACES environment variable has not been explicitly set and thus defaults to the OpenMP cores setting which defines an OpenMP place as a single processor core. Together with the clause num_threads(ncores), the number of OpenMP threads is equal to the number of OpenMP places, i.e. to the number of processor cores. We did not measure any performance improvements for a higher thread count.

The proc_bind(spread) clause additionally binds each OpenMP thread to one OpenMP place which lowers inter-node or inter-socket communication and improves local memory access. Moreover, with the spread thread affinity policy, consecutive OpenMP threads are spread across OpenMP places which can be beneficial if the user decides to set ncores smaller than the number of processor cores.

5.3. Tensor Shapes

We have used asymmetrically and symmetrically shaped tensors in order to cover many use cases. The dimension tuples of both shape types are organized within two three-dimensional arrays with which tensors are initialized. The dimension array for the first shape type contains $720 = 9 \times 8 \times 10$ dimension tuples where the row number is the tensor order ranging from 2 to 10. For each tensor order, 8 tensor instances with increasing tensor size is generated. A special feature of this test set is that the contraction dimension and the leading dimension are disproportionately large. The second set consists of $336 = 6 \times 8 \times 7$ dimensions tuples where the tensor order ranges from 2 to 7 and has 8 dimension tuples for each order. Each tensor dimension within the second set is 2^{12}, 2^8, 2^6, 2^5, 2^4 and 2^3. A detailed explanation of the tensor shape setup is given in [13, 17]. If not otherwise mentioned, both tensors $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ are stored according to the first-order tensor layout.

6. Results and Discussion

6.1. Slicing Methods

This section analyzes the performance of the two proposed slicing methods `<slice>` and `<subtensor>` that have been discussed in section 4.4. Figure 1 contains eight performance contour plots of four `ttm` functions `<par-loop>` and `<par-gemm>` that either compute the slice-matrix product with subtensors `<subtensor>` or tensor slices `<slice>`. Each contour level within the plots represents a mean GFLOPS/core value that is averaged across tensor sizes.

Every contour plot contains all applicable TTM cases listed in Table 1. The first column of performance values is generated by `gemm` belonging to the TTM case 3, except the first element which corresponds to TTM case 2. The first row, excluding the first element, is generated by TTM case 6 function. TTM case 7 is covered by the diagonal line of performance values when $q = p$. Although Figure 1 suggests that $q > p$ is possible, our profiling program ensures that $q = p$. TTM case 8 with multiple `gemm` calls is represented by the triangular region which is defined by $1 < q < p$.

Function `<par-loop>` with `<slice>` runs on average with 34.96 GFLOPS/core (1.67 TFLOPS) with asymmetrically shaped tensors. With a maximum performance of 57.805 GFLOPS/core (2.77 TFLOPS), it performs on average 89.64% faster than function `<par-loop>` with `<subtensor>`. The slowdown with subtensors at $q = p-1$ or $q = p-2$ can be explained by the small loop count of the function that are 2 and 4, respectively. While function `<par-loop>` with tensor slices is affected by the tensor shapes for dimensions $p = 3$ and $p = 4$ as well, its performance improves with increasing order due to the increasing loop count. The performance drops and their corresponding locations on the performance plots have also been mentioned in [1].

Function `<par-loop>` with tensor slices achieves on average 17.34 GFLOPS/core (832.42 GFLOPS) with symmetrically shaped tensors. In this case, `<par-loop>` with subtensors achieves a mean throughput of 17.62 GFLOPS/core (846.16 GFLOPS) and is on average 9.89% faster than the `<slice>` version. The performances of both functions are monotonically decreasing with increasing tensor order, see plots (1.c) and (1.d) in Figure 1. The average performance decrease of both functions can be approximated by a cubic polynomial with the coefficients -35, 640, -3848 and 8011. The decreasing performance behavior for symmetrically shaped tensors has also been described in [1].

Function `<par-gemm>` with tensor slices averages 36.42 GFLOPS/core (1.74 TFLOPS) and achieves up to 57.91 GFLOPS/core (2.77 TFLOPS) with asymmetrically shaped tensors. With subtensors, function `<par-gemm>` exhibits almost identical performance characteristics and is on average only 3.42% slower than its counterpart with tensor slices.

For symmetrically shaped tensors, `<par-gemm>` with subtensors and tensor slices achieve a mean throughput 15.98 GFLOPS/core (767.31 GFLOPS) and 15.43 GFLOPS/core (740.67 GFLOPS), respectively. However, function `<par-gemm>` with `<subtensor>` is on average 87.74% faster than the `slice` which is hardly visible due to small performance values around 5 GFLOPS/core or less whenever $q < p$ and the dimensions are smaller than 256. The speedup of the `<subtensor>` version can be explained by the smaller loop count and slice-matrix multiplications with larger tensor slices.

6.2. Parallelization Methods

This section discusses the performance results of the two parallelization methods `<par-gemm>` and `<par-loop>` using the same Figure 1.

With asymmetrically shaped tensors, both `<par-gemm>` functions with subtensors and tensor slices compute the tensor-matrix product on average with 36 GFLOPS/core and outperform function `<par-loop>` with `<subtensor>` on average by a factor of 2.31. The speedup can be explained by the performance drop of function `<par-loop>` `<subtensor>` to 3.49 GFLOPS/core at $q = p - 1$ while both `<par-gemm>` functions operate around 39 GFLOPS/core. Function `<par-loop>` with tensor slices performs better for reasons explained in the previous subsection. It is on average 30.57% slower than its `<par-gemm>` version due to the aforementioned performance drops.

In case of symmetrically shaped tensors, `<par-loop>` with subtensors and tensor slices outperform their corresponding `<par-gemm>` counterparts by 23.3% and 32.9%, respectively. The speedup mostly occurs when $1 < q < p$ where the performance gain is a factor of 2.23. This performance behavior can be expected as the tensor slice sizes decreases for the eighth case with increasing tensor order causing the parallel slice-matrix multiplication to perform on smaller matrices. In contrast, `<par-loop>` can execute small single-threaded slice-matrix multiplications in parallel.

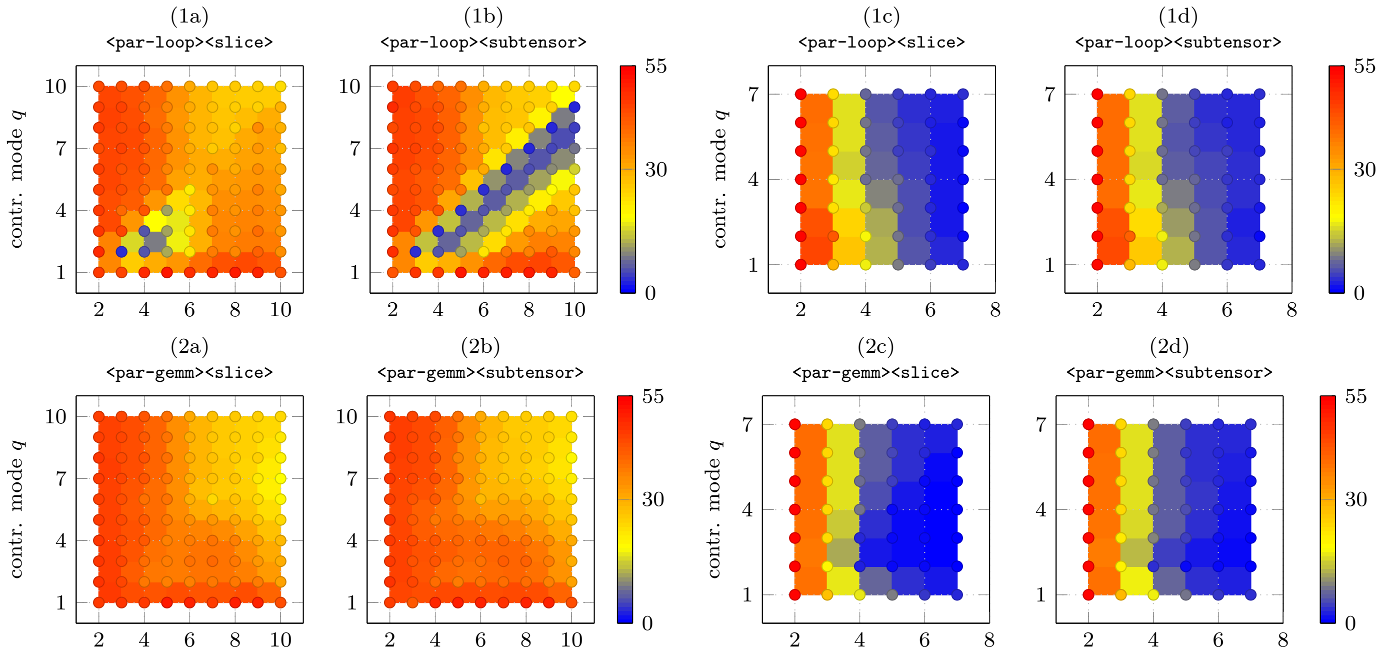

Figure 1: Performance contour plots in double-precision GFLOPS/core of the proposed TTM algorithms `<par-loop>` and `<par-gemm>` with varying tensor orders p and contraction modes q. The top row of maps (1x) depict measurements of the `<par-loop>` versions while the bottom row of maps with number (2x) contain measurements of the `<par-gemm>` versions. Tensors are asymmetrically shaped on the left four maps (a,b) and symmetrically shaped on the right four maps (c,d). Tensor $\underline{\mathbf{A}}$ and $\underline{\mathbf{C}}$ have the first-order while matrix $\mathbf{B}$ has the row-major ordering. All functions have been measured on an Intel Xeon Gold 5318Y.

6.3. *Loops Over Gemm*

The contour plots in Figure 1 contain performance data that are generated by all applicable TTM cases of each `ttm` function. Yet, the presented slicing or parallelization methods only affect the eighth case, while all other TTM cases apply a single multi-threaded `gemm`. The following analysis will consider performance values of the eighth case in order to have a more fine grained visualization and discussion of the loops over `gemm` implementations. Figure 2 contains cumulative performance distributions of all the proposed algorithms including the `<mkl-batch-gemm>` and `<combined>` functions for case 8 only. Moreover, the experiments have been additionally executed on the AMD EPYC processor and with the column-major ordering of the input matrix as well.

The probability x of a point (x, y) of a distribution function for a given algorithm corresponds to the number of test instances for which that algorithm that achieves a throughput of either y or less. For instance, function `<mkl-batch-gemm>` computes the tensor-matrix product with asymmetrically shaped tensors in 25% of the tensor instances with equal to or less than 10 GFLOPS/core. Please note that the four plots on the right, plots (c) and (d), have a logarithmic y-axis for a better visualization.

6.3.1. *Combined Algorithm and Batched GEMM*

Given a row-major matrix ordering, the combined function `<combined>` achieves on the Intel processor a median throughput of 36.15 and 4.28 GFLOPS/core with asymmetrically and symmetrically shaped tensors. Reaching up to 46.96 and 45.68 GFLOPS/core, it is on par with `<par-gemm>` with subtensors and `<par-loop>` with tensor slices and outperforms them for some tensor instances. Note that both functions run significantly slower either with asymmetrically or symmetrically shaped tensors. The observable superior performance distribution of `<combined>` can be attributed to the heuristic which switches between `<par-loop>` and `<par-gemm>` depending on the inner and outer loop count.

Function `<mkl-batch-gemm>` of the BLAS-like extension library has a performance distribution that is akin to the `<par-loop>` with subtensors. In case of asymmetrically shaped tensors, all functions except `<par-loop>` with subtensors outperform `<mkl-batch-gemm>` on average by a factor of 2.57 and up to a factor 4 for $2 \leq q \leq 5$ with $q+2 \leq p \leq q+5$. In contrast, `<par-loop>` with subtensors and `<mkl-batch-gemm>` show a similar performance behavior in the plot (1c) and (1d) for symmetrically shaped tensors, running on average 3.55 and 8.38 times faster than `<par-gemm>` with subtensors and tensor slices, respectively. Function `<par-loop>` with tensor slices underperforms for $p > 3$, i.e. when the tensor dimensions are less than 64. These observations have also been mentioned in [1].

6.3.2. *Matrix Formats*

The cumulative performance distributions in Figure 2 suggest that the storage format of the input matrix has only a minor impact on the performance. The Euclidean distance between normalized row-major and column-major performance values is around 5 or less with a maximum

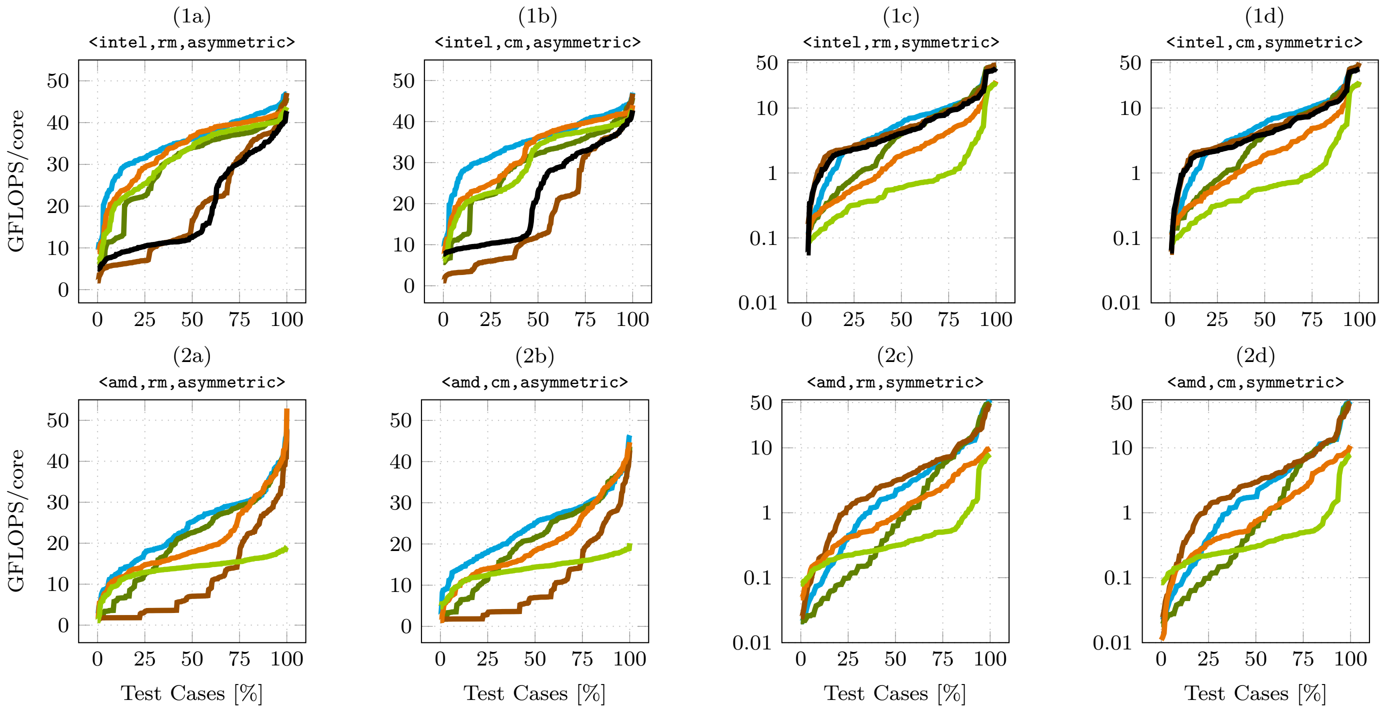

Figure 2: Cumulative performance distributions in double-precision GFLOPS/core of the proposed algorithms for the eighth case. Each distribution belongs to one algorithm: `<mkl-batch-gemm>` (black line), `<combined>` (blue line), `<par-gemm>` (light green line) and `<par-loop>` (dark green line) using tensor slices, `<par-gemm>` (orange line) and `<par-loop>` (brown line) using subtensors. The top row of maps (1x) depict measurements performed on an Intel Xeon Gold 5318Y with the MKL while the bottom row of maps with number (2x) contain measurements performed on an AMD EPYC 9354 with the AOCL. Tensors are asymmetrically shaped in (a) and (b) and symmetrically shaped in (c) and (d). Input matrix has the row-major ordering (rm) in (a) and (c) and column-major ordering (cm) in (b) and (d).

dissimilarity of 11.61 or 16.97, indicating a moderate similarity between the corresponding row-major and column-major data sets. Moreover, their respective median values with their first and third quartiles differ by less than 5% with three exceptions where the difference of the median values is between 10% and 15%.

6.3.3. BLAS Libraries

This subsection compares the performance of functions that use Intel's Math Kernel Library (MKL) on the Intel Xeon Gold 5318Y processor with those that use the AMD Optimizing CPU Libraries (AOCL) on the AMD EPYC 9354 processor. Limiting the performance evaluation to the eighth case, MKL-based functions with asymmetrically shaped tensors run on average between 1.48 and 2.43 times faster than those with the AOCL. For symmetrically shaped tensors, MKL-based functions are between 1.93 and 5.21 times faster than those with the AOCL. In general, MKL-based functions achieve a speedup of at least 1.76 and 1.71 compared to their AOCL-based counterpart when asymmetrically and symmetrically shaped tensors are used.

6.4. Layout-Oblivious Algorithms

Figure 3 contains four subfigures with box plots summarizing the performance distribution of the `<combined>` function using the AOCL and MKL. Every kth box plot has been computed from benchmark data with symmetrically shaped order-7 tensors that has a k-order tensor layout. The 1-order and 7-order layout, for instance, are the first-order and last-order storage formats of an order-7 tensor[2]. Note that `<combined>` only calls `<par-loop>` with subtensors.

The reduced performance of around 1 and 2 GFLOPS can be attributed to the fact that contraction and leading dimensions of symmetrically shaped subtensors are at most 48 and 8, respectively. When `<combined>` is used with MKL, the relative standard deviations (RSD) of its median performances are 2.51% and 0.74%, with respect to the row-major and column-major formats. The RSD of its respective interquartile ranges (IQR) are 4.29% and 6.9%, indicating a similar performance distributions. Using `<combined>` with AOCL, the RSD of its median performances for the row-major and column-major formats are 25.62% and 20.66%, respectively. The RSD of its respective IQRs are 10.83% and 4.31%, indicating a similar performance distributions.

A similar performance behavior can be observed also for other `ttm` variants such as `par-loop` with tensor slices or `par-gemm`. The runtime results demonstrate that the function performances stay within an acceptable range independent for different k-order tensor layouts and show

[2] The k-order tensor layout definition is given in section 3.4

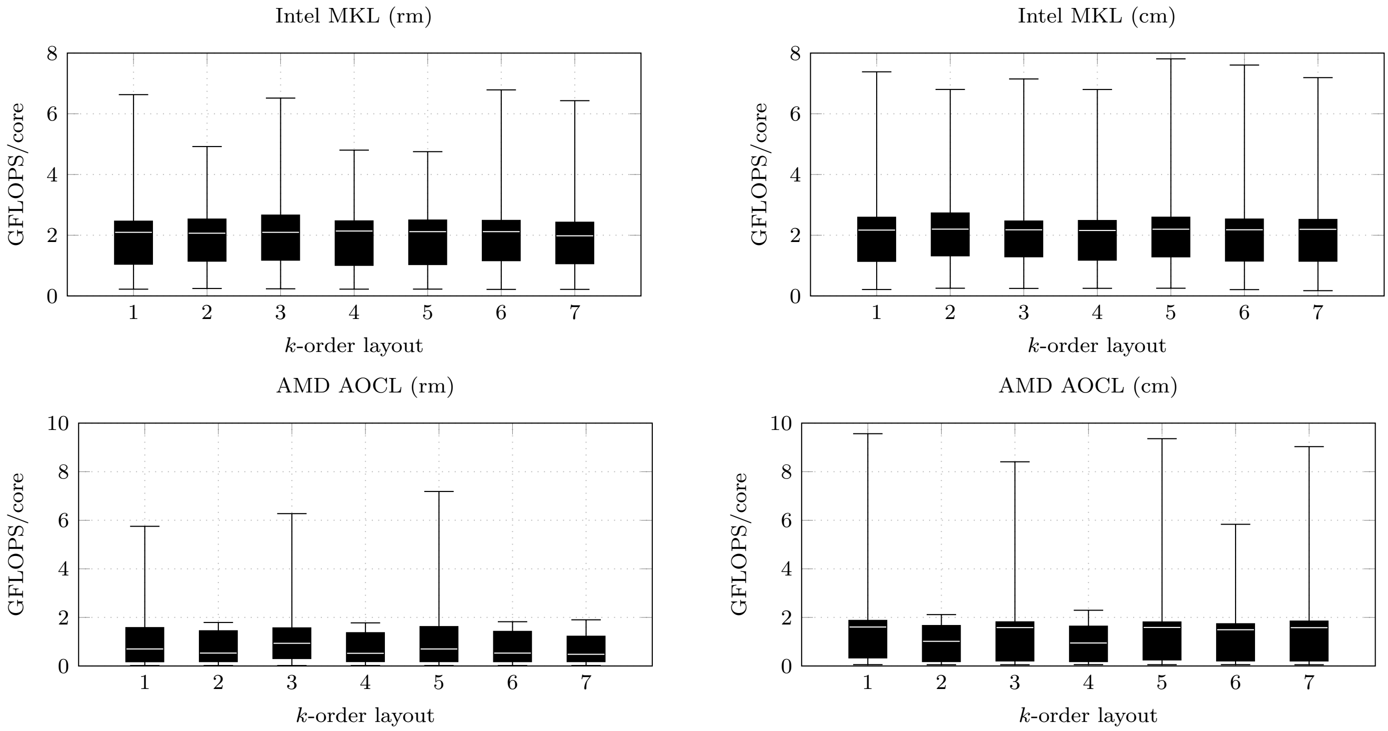

Figure 3: Box plots visualizing performance statics in double-precision GFLOPS/core of the function with row-major (left) or column-major matrices (right). Box plot number k denotes the k-order tensor layout of symmetrically shaped tensors with order 7.

that our proposed algorithms are not designed for a specific tensor layout.

6.5. Other Approaches

This subsection compares our best performing algorithm with libraries that do not use the LoG approach. **TCL** implements the TTGT approach with a high-perform tensor-transpose library **HPTT** which is discussed in [8]. **TBLIS** (v1.2.0) implements the GETT approach that is akin to BLIS' algorithm design for the matrix multiplication [9]. The tensor extension of **Eigen** (v3.4.9) is used by the Tensorflow framework. Library **LibTorch** (v2.4.0) is the C++ distribution of PyTorch [15]. **TLIB** denotes our library which only calls the previously presented algorithm <combined>. We will use performance or percentage tuples of the form (TCL, TBLIS, LibTorch, Eigen) where each tuple element denotes the performance or runtime percentage of a particular library.

Figure 2 compares the performance distribution of our implementation with the previously mentioned libraries. Using MKL on the Intel CPU, our implementation (TLIB) achieves a median performance of 38.21 GFLOPS/core (1.83 TFLOPS) and reaches a maximum performance of 51.65 GFLOPS/core (2.47 TFLOPS) with asymmetrically shaped tensors. It outperforms the competing libraries for almost every tensor instance within the test set. The median library performances are (24.16, 29.85, 28.66, 14.86) GFLOPS/core reaching on average (84.68, 80.61, 78.00, 36.94) percent of TLIB's throughputs. In case of symmetrically shaped tensors other libraries on the right plot in Figure 2 run at least 2 times slower than TLIB except for TBLIS. TLIB's median performance is 8.99 GFLOPS/core, other libraries achieve a median performances of (2.70, 9.84, 3.52, 3.80) GFLOPS/core. On average their performances constitute (44.65, 98.63, 53.32, 31.59) percent of TLIB's throughputs.

On the AMD CPU, our implementation with AOCL computes the tensor-times-matrix product on average with 24.28 GFLOPS/core (1.55 TFLOPS) and reaches a maximum performance of 45.84 GFLOPS/core (2.93 TFLOPS) with asymmetrically shaped tensors. TBLIS reaches 26.81 GFLOPS/core (1.71 TFLOPS) and is slightly faster than TLIB. However, TLIB's upper performance quartile with 30.82 GFLOPS/core is slightly larger. TLIB outperforms other competing libraries that have a median performance of (8.07, 16.04, 11.49) GFLOPS/core reaching on average (27.97, 62.97, 54.64) percent TLIB's throughputs. In case of symmetrically shaped tensors, TLIB outperforms all other libraries with 7.52 GFLOPS/core (481.39 GFLOPS) and a maximum performance of 47.78 GFLOPS/core (3.05 TFLOPS). Other libraries perform with (2.03, 6.18, 2.64, 5.58) GFLOPS/core and reach (44.94, 86.67, 57.33, 69.72) percent of TLIB's throughputs. We have observed that TCL and LibTorch have a median performance of less than 2 GFLOPS/core in the 3rd and 8th TTM case which is less than 6% and 10% of TLIB's median performance with asymmetrically and symmetrically shaped tensors, respectively.

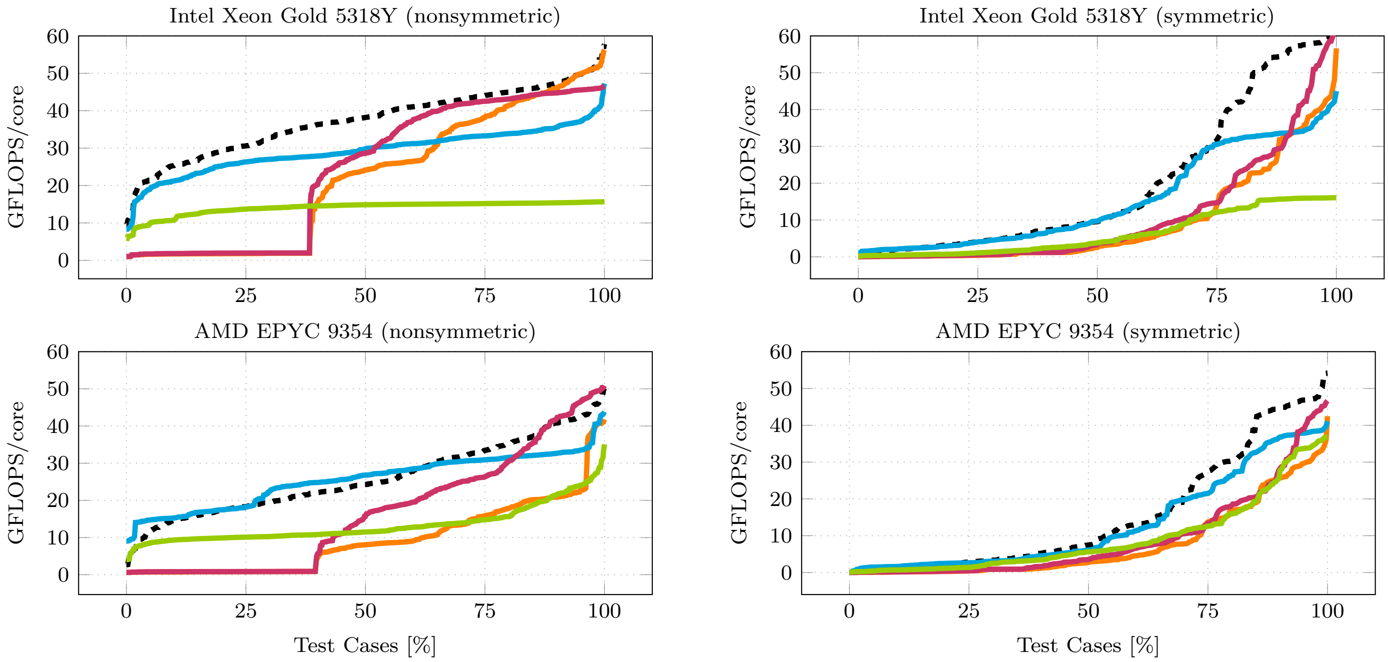

Figure 4: Cumulative performance distributions of TTM implementations in double-precision GFLOPS/core. Each distribution corresponds to a library: **TLIB**[ours] ( ), **TCL** ( ), **TBLIS** ( ), **LibTorch** ( ), **Eigen** ( ). Libraries have been tested with asymmetrically-shaped (left plot) and symmetrically-shaped tensors (right plot).

While all libraries run on average 25% slower than TLIB across all TTM cases, there are few exceptions. On the AMD CPU, TBLIS reaches 101% of TLIB's performance for the 6th TTM case and LibTorch performs as fast as TLIB for the 7th TTM case for asymmetrically shaped tensors. One unexpected finding is that LibTorch achieves 96% of TLIB's performance with asymmetrically shaped tensors and only 28% in case of symmetrically shaped tensors.

On the Intel CPU, LibTorch is on average 9.63% faster than TLIB in the 7th TTM case. The TCL library runs on average as fast as TLIB in the 6th and 7th TTM cases . The performances of TLIB and TBLIS are in the 8th TTM case almost on par, TLIB running about 7.86% faster. In case of symmetrically shaped tensors, all libraries except Eigen outperform TLIB by about 13%, 42% and 65% in the 7th TTM case. TBLIS and TLIB perform equally well in the 8th TTM case, while other libraries only reach on average 30% of TLIB's performance.

7. Conclusion and Future Work

We have presented efficient layout-oblivious algorithms for the compute-bound tensor-matrix multiplication that is essential for many tensor methods. Our approach is based on the LOG-method and computes the tensor-matrix product in-place without transposing tensors. It applies the flexible approach described in [13] and generalizes the findings on tensor slicing in [11] for linear tensor layouts. The resulting algorithms are able to process dense tensors with arbitrary tensor order, dimensions and with any linear tensor layout all of which can be runtime variable.

The base algorithm has been divided into eight different TTM cases where seven of them perform a single `cblas_gemm`. We have presented multiple algorithm variants for the general TTM case which either calls a single- or multi-threaded `cblas_gemm` with small or large tensor slices in parallel or sequentially. We have developed a simple heuristic that selects one of the variants based on the performance evaluation in the original work [1]. With a large set of tensor instances of different shapes, we have evaluated the proposed variants on an Intel Xeon Gold 5318Y and an AMD EPYC 9354 CPUs.

Our performance tests show that our algorithms are layout-oblivious and do not need layout-specific optimizations, even for different storage ordering of the input matrix. Despite the flexible design, our best-performing algorithm is able to outperform Intel's BLAS-like extension function `cblas_gemm_batch` by a factor of 2.57 in case of asymmetrically shaped tensors. Moreover, the presented performance results show that TLIB is able to compute the tensor-matrix product on average 25% faster than other state-of-the-art implementations for a majority of tensor instances.

Our findings show that the LoG-based approach is a viable solution for the general tensor-matrix multiplication which can be as fast as or even outperform efficient GETT-based implementations. Hence, other actively developed libraries such as LibTorch and Eigen might benefit from implementing the proposed algorithms. Our headeronly library provides C++ interfaces and a python module which allows frameworks to easily integrate our library.

In the near future, we intend to incorporate our implementations in TensorLy, a widely-used framework for tensor computations [18, 19]. Using the insights provided in [11] could help to further increase the performance. Additionally, we want to explore to what extend our approach can be applied for the general tensor contractions.

7.0.1. Source Code Availability

Project description and source code can be found at `https://github.com/bassoy/ttm`. The sequential tensor-matrix multiplication of TLIB is part of Boost's uBLAS library.

References

[1] C. S. Başsoy, Fast and layout-oblivious tensor-matrix multiplication with blas, in: International Conference on Computational Science, Springer, 2024, pp. 256–271.

[2] E. Karahan, P. A. Rojas-López, M. L. Bringas-Vega, P. A. Valdés-Hernández, P. A. Valdes-Sosa, Tensor analysis and fusion of multimodal brain images, Proceedings of the IEEE 103 (9) (2015) 1531–1559.

[3] E. E. Papalexakis, C. Faloutsos, N. D. Sidiropoulos, Tensors for data mining and data fusion: Models, applications, and scalable algorithms, ACM Transactions on Intelligent Systems and Technology (TIST) 8 (2) (2017) 16.

[4] N. Lee, A. Cichocki, Fundamental tensor operations for large-scale data analysis using tensor network formats, Multidimensional Systems and Signal Processing 29 (3) (2018) 921–960.

[5] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM review 51 (3) (2009) 455–500.

[6] B. W. Bader, T. G. Kolda, Algorithm 862: Matlab tensor classes for fast algorithm prototyping, ACM Trans. Math. Softw. 32 (2006) 635–653.

[7] E. Solomonik, D. Matthews, J. Hammond, J. Demmel, Cyclops tensor framework: Reducing communication and eliminating load imbalance in massively parallel contractions, in: Parallel & Distributed Processing (IPDPS), 2013 IEEE 27th International Symposium on, IEEE, 2013, pp. 813–824.

[8] P. Springer, P. Bientinesi, Design of a high-performance gemm-like tensor–tensor multiplication, ACM Transactions on Mathematical Software (TOMS) 44 (3) (2018) 28.

[9] D. A. Matthews, High-performance tensor contraction without transposition, SIAM Journal on Scientific Computing 40 (1) (2018) C1–C24.

[10] E. D. Napoli, D. Fabregat-Traver, G. Quintana-Ortí, P. Bientinesi, Towards an efficient use of the blas library for multilinear tensor contractions, Applied Mathematics and Computation 235 (2014) 454 – 468.

[11] J. Li, C. Battaglino, I. Perros, J. Sun, R. Vuduc, An input-adaptive and in-place approach to dense tensor-times-matrix multiply, in: High Performance Computing, Networking, Storage and Analysis, 2015, IEEE, 2015, pp. 1–12.

[12] Y. Shi, U. N. Niranjan, A. Anandkumar, C. Cecka, Tensor contractions with extended blas kernels on cpu and gpu, in: 2016 IEEE 23rd International Conference on High Performance Computing (HiPC), 2016, pp. 193–202.

[13] C. Bassoy, Design of a high-performance tensor-vector multiplication with blas, in: International Conference on Computational Science, Springer, 2019, pp. 32–45.

[14] F. Pawlowski, B. Uçar, A.-J. Yzelman, A multi-dimensional morton-ordered block storage for mode-oblivious tensor computations, Journal of Computational Science 33 (2019) 34–44.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019).

[16] L.-H. Lim, Tensors and hypermatrices, in: L. Hogben (Ed.), Handbook of Linear Algebra, 2nd Edition, Chapman and Hall, 2017.

[17] C. Bassoy, V. Schatz, Fast higher-order functions for tensor calculus with tensors and subtensors, in: International Conference on Computational Science, Springer, 2018, pp. 639–652.

[18] J. Cohen, C. Bassoy, L. Mitchell, Ttv in tensorly, Tensor Computations: Applications and Optimization (2022) 11.

[19] J. Kossaifi, Y. Panagakis, A. Anandkumar, M. Pantic, Tensorly: Tensor learning in python, Journal of Machine Learning Research 20 (26) (2019) 1–6.