

PASTA: A Parallel Sparse Tensor Algorithm Benchmark Suite

JIAJIA LI, Pacific Northwest National Laboratory

YUCHEN MA, Hangzhou Dianzi University

XIAOLONG WU, Virginia Tech

ANG LI, Pacific Northwest National Laboratory

KEVIN BARKER, Pacific Northwest National Laboratory

Tensor methods have gained increasingly attention from various applications, including machine learning, quantum chemistry, healthcare analytics, social network analysis, data mining, and signal processing, to name a few. Sparse tensors and their algorithms become critical to further improve the performance of these methods and enhance the interpretability of their output. This work presents a sparse tensor algorithm benchmark suite (PASTA) for single- and multi-core CPUs. To the best of our knowledge, this is the first benchmark suite for sparse tensor world. PASTA targets on: 1) helping application users to evaluate different computer systems using its representative computational workloads; 2) providing insights to better utilize existed computer architecture and systems and inspiration for the future design. This benchmark suite will be publicly released.

ACM Reference format:

Jiajia Li, Yuchen Ma, Xiaolong Wu, Ang Li, and Kevin Barker. 2016. PASTA: A Parallel Sparse Tensor Algorithm Benchmark Suite. 1, 1, Article 1 (January 2016), 25 pages.

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Tensors draw increasing attention from various domains, such as machine learning, quantum chemistry, healthcare analytics, social network analysis, data mining, and signal processing, to name a few. Tensor methods have been noted for their ability to discover multi-dimensional inherent relationships from underlying application logic. A tensor is a multi-dimensional array, generalized matrices and vectors to more dimensions. In data-oriented tensor applications [22, 50, 54, 106, 121], sparse tensors are often found, where most of its entries are zeros.

High-performance computing (HPC) now enters the era of extreme heterogeneity. As many general purpose accelerators, such as Graphics Processing Unit (GPUs), Intel Xeon Phi, and Field-Programmable Gate Array (FPGAs), and domain-specific architectures, such as near-memory, thread migratory architecture Emu [49] and Google Tensor processing unit (TPU) [63], emerge, it is natural to ask whether the critical sparse-tensor based algorithms can be efficiently executed on these platforms, with their non-regular parallelism to be effectively exploited. However, the lack of a concrete, comprehensive, and easy to use sparse tensor algorithm benchmark suite prevents us from answering this question easily.

In this paper, we fill this gap by proposing a PArallel Sparse Tensor Algorithm benchmark suite called PASTA. PASTA incorporates various sparse tensor algorithms and operations, serving as a handy tool for application developers to assess different platforms, in terms of their tensor

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2016 ACM. XXXX-XXXX/2016/1-ART1 \$15.00

DOI: 10.1145/nnnnnnn.nnnnnnn

processing capability. Consisting state-of-the-art sequential and parallel versions, while adopting the most popular sparse tensor format COO, PASTA can also supply a fair baseline for evaluating performance improvement brought by new sparse tensor methods. Application developers seeking to exploit tensor sparsity for further performance speedup may also find it useful as a good reference.

This paper makes the following contributions:

- We show the importance of sparse tensor operations and tensor methods in diverse tensor applications. (Section 3)
- We extract 12 computational sparse tensor operations as PASTA workloads: Tensor Element-Wise operations – T_{EW} -eq (addition/subtraction/multiplication/division) and T_{EW} (addition/subtraction/multiplication), Tensor-Scalar operations – T_s addition/multiplication, Tensor-Times-Vector operation (T_{TV}), Tensor-Times-Matrix operation (T_{TM}), and Matricized Tensor Times Khatri-Rao Product ($MTTKRP$). (Section 4)
- We implement sequential and multicore parallel algorithms for all workloads, based on the most popular coordinate (COO) sparse tensor format. Our experiments and analysis show the usefulness of PASTA on single- and multi-core CPUs. (Section 5, 6, 7)

2 MOTIVATION

This work is motivated by first demonstrating the challenges of sparse tensor algorithms and then illustrating that existed libraries or toolsets cannot meet the requirements of a benchmark suite from diversity, timeliness, research support, and dataset four aspects.

2.1 Challenges of Sparse Tensor Algorithms

We summarize the challenges of sparse tensor algorithms into five points:

The curse of dimensionality refers to the issue that the number of entries of an intermediate or output tensor can grow exponentially with the tensor order, resulting in significant computational and storage overheads. Even when the tensor is structurally sparse, meaning it consists mostly of zero entries, the execution time of one important tensor method, CP decomposition introduced in Section 3.1, generally grows quadratically with the number of non-zeros [7, 8]. And there is an increasing interest in applications involving a large number of dimensions [34, 79, 97], which makes this problem more difficult.

Mode orientation refers to the issue of a particular storage format favoring the iteration of tensor modes in a certain sequence, which is of particular concern in the sparse case. Since most methods of interest require more than one sequence, being efficient for every sequence generally requires storing the tensor in multiple formats, thereby trading extra memory for speed. A question arises, that is whether one can achieve both a neutral mode orientation and compact storage which also helps reduce memory footprint.

Tensor transformation(s) refers to a common pattern for attaining speed in some implementations of tensor algorithms, which starts by reorganizing the tensor into a matrix and then perform equivalent matrix operations using highly tuned linear algebra libraries. Done naïvely, this approach appears to require an extra memory copy, which can even come to dominate the overall running time. We observe instances in which such a copy consumes 70% or more of the total running time (in the case of a T_{TM} operation).

Irregularity refers to two issues. The first is that a tensor may have dimension sizes that vary widely; the second is that a sparse tensor may have an irregular non-zero pattern, resulting in irregular memory references.

Arbitrary tensor orders generate various implementations of a tensor operation. For the sake of performance, programmers usually implement and optimize third-order tensor algorithms apart

from higher-order ones. These implementations makes no one optimization method can fit all variations, e.g., different number of loops and diverse memory access behavior.

These challenges bring non-trivial computational and storage overheads, and some of them are even harder to overcome than their counterparts in classical linear algebra. To overcome these challenges, it is necessary to build a sparse tensor benchmark suite to evaluate diverse algorithms and computer systems.

2.2 Requirements for a Benchmark Suite

By surveying some benchmark suites [12, 20, 35, 73, 80, 111, 151], we present the following four requirements for a benchmark suite.

Diversity. We analyze diversity from two aspects: *application diversity* and *platform diversity*. *Application diversity* means a benchmark suite should represent a broad and representative applications. For example, EEMBC benchmark suite [111] is developed for autonomous driving, mobile imaging, the Internet of Things, mobile devices, and many other applications; PARSEC benchmark suite [12] covers computer vision, video encoding, financial analytics, animation physics and image processing, etc.. Sparse tensor methods have a broad application domains (refer to Section 3.2), the workloads in our benchmark suite also need to represent the diversity of these domains. *Platform diversity* is that a benchmark suite should support different computer architectures and platforms, especially the emerging ones. For example, SPEC benchmarks [35] supports scientific applications on diverse platforms: CPUs, distributed platforms, accelerators, web servers, cloud platforms, etc. A recent Tartan benchmark [82] collected kernels from machine learning, data analysis, high performance simulation, molecular dynamics and so on and optimized them on multi-GPU platforms.

Timeliness. A benchmark suite should be kept updated by including the state-of-the-art *data structures*, *algorithms*, and *optimization techniques*. Especially for sparse data, the data structure is closely relevant to the performance of its algorithm. This phenomenon has been observed from sparse matrices, where different sparse formats behave quite differently on diverse input matrices [87, 119, 139, 163]. As mentioned in the work [12], an outdated algorithm cannot well reflect the current status of an application. This can easily mislead the researchers using this benchmark suite to test a machine's behavior. As the computer architectures keep evolving, an under-optimized code, e.g., sequential benchmark programs for a multicore machine, cannot be a fair measurement. Optimized implementations for architectures have to be taken account.

Research support. Research support also includes two aspects: support of domain research and benchmarked workload research. The former requires a benchmark suite to be *compatible*, while the latter requires it to be *extensible*. Since some workloads are still open research problems in an application domain, a compatible workload should be able to do easy comparison with other research work by supporting unified input/output format and interface to high-level applications. The workload research mainly develops its high performance, power or other efficiency. An extensible workload is easy to be assembled with new data structures, algorithms, and optimization techniques.

Dataset. Data becomes essential to data-intensive applications and their workloads which widely exist in real world. Traditionally, two types of dataset are considered: *synthetic* and *real* data. Real data comes directly from real-world applications, which can best reflects the application features. However, due to some factors such as information protection, sensitive data, etc., researchers are usually short of data. Thus, synthetic data are generated according to some regulations and scenarios from applications.

Domains	Tensor Methods	Workloads
Machine Learning	CPD, TPM, Tucker, TT, hTucker	Ts, MTTKRP, TTV, TTM, TTT
Healthcare Analytics	CPD	MTTKRP
Social Network Analysis	CPD, Tucker	TTM
Quantum Chemistry	CPD, Tucker	Ts, TEW, TTM, MTTKRP, TTT
Brain Signal Analysis	CPD	MTTKRP
Personalized web search	CPD, Tucker	MTTKRP, TTM
Recommendation systems	CPD, Tucker	MTTKRP, TTM
Signal Processing	CPD	MTTKRP
Direct Numerical Simulation	Tucker	TTM
Power Grid	CPD, Tucker	MTTKRP, TTM

Table 1. The relationship between tensor domains, tensor methods, and workloads.

2.3 PASTA in Need

Some tensor libraries or toolsets have existed for sparse tensor algorithms. The most popular libraries are Tensor Toolbox [8] and TensorLab [147]. They are both implemented using MATLAB. The main shortcoming is that these two libraries are hard to be implemented on various platforms, such as multicore CPUs and GPUs, which violates the platform diversity requirement. Besides, their performance efficiency is low because of MATLAB environment. Recently, many other highly performance efficient libraries emerge, such as SPLATT [130], Cyclops Tensor Framework (CTF) [132], DFacTo [24], GigaTensor [65], HyperTensor [69], GenTen [110], to name a few. However, these libraries are specific to one or two particular sparse tensor operations, this violates the application diversity requirement. Beyond these, the requirements of timeliness, research support, and dataset are barely met by these libraries. Our PASTA is proposed to meet all the requirements from our continuous effort.

3 TENSOR METHODS AND APPLICATIONS

This section describes the broad applications of tensors methods in diverse domains, along with the tensor methods and their computational operations. The summarized form is presented in Table 1.

3.1 Tensor Methods

In this section, we summarize tensor methods in three categories: tensor decompositions, tensor network models, and tensor regression. Though tensor network models also belong to tensor decomposition methods, because of their network format and more emphasizing on high-order tensors, we discuss them separately.

3.1.1 Tensor Decompositions. We introduce three low-rank tensor decompositions which have applications for sparse data.

CPD. The CP decomposition (CPD) was first introduced in 1927 by Hitchcock [51], and independently introduced by others [18, 47]. CPD decomposes an N th-order tensor into a sum of component rank-one tensors with different weights [74]. In a low-rank approximation, a tensor rank R is chosen to be a small number less than 100. From a data science standpoint, the results can be interpreted by viewing the tensor as being composed of R latent rank-1 factors. CPD has proven both scalable and effective in many applications in Section 3.2.

Other variants of CPD exist by restructuring of the factors or their constraints to accommodate diverse situations, such as INDSCAL [18], CANDELINC [19], PARAFAC2 [48, 109], and DEDICOM [47]. Many CPD methods have been proposed in a broad area of research, such as Alternating

Least Squares (ALS) based methods [47, 68, 69, 74], block coordinate descent (BCD) based methods [88, 93], Gradient Descent based methods [11, 113, 128, 138], quasi-Newton and Nonlinear Least Squares (NLS) based methods [22, 45, 57, 117, 138, 145, 154], alternating optimization (AO) with the alternating direction method of multipliers (ADMM) based methods [13, 124], exact line search based methods [112, 137], and randomized/sketching methods [9, 21, 104, 115, 136, 148]. Sparse CPD comes from two aspects: the sparse tensor from applications [7, 22–24, 65, 70, 74, 83, 84, 86, 89, 110, 113, 121, 126, 129, 130] and the constrained sparse factors from some CPD models [50, 54, 106].

The computational bottleneck of CPD is the matriced tensor-times-Khatri-Rao product (MTTKRP) (will be described in Section 4.6).

Tucker. Tucker decomposition, first introduced by Ledyard R. Tucker [146], provides a more general decomposition. It decomposes an N th-order tensor into a small-sized N th-order core tensor along with N factor matrices that are all orthogonal. The core tensor models a potentially complex pattern of mutual interaction between tensor modes. Its size determined by N ranks which can be chosen according to the work [72]. In a low-rank approximation, the rank sizes are usually less than 100.

Some variants of Tucker decomposition are PARATUCK2 [46], lossy Tucker decomposition [164], and so on. Methods for Tucker decomposition include higher-order SVD (HOSVD) [32], truncated HOSVD [32], Alternating Least Squares (ALS) based methods [66], the popular higher-order orthogonal iteration (HOOI) [33], Newtonfi?!Grassmann optimization [36]. Sparse Tucker also comes from two aspects: the sparse tensor from applications [83, 89, 90, 127] and the constrained sparse factors.

The computational tensor kernel of Tucker decomposition is the Tensor-Times-Matrix operation (T_{TM}) (will be described in Section 4.4).

TPM. Tensor power method [5, 33] is an approach for orthogonal tensor decomposition, which decomposes a symmetric tensor into a collection of orthogonal vectors with corresponding positive scalars as weights. Some variations have been proposed [5, 158]. When the tensor is sparse, we need to use sparse method correspondingly.

The computational tensor kernel of tensor power method is the Tensor-Times-Vector operation (T_{TV}) (will be described in Section 4.3).

3.1.2 TENSOR NETWORK MODELS. CPD and Tucker decompositions assume a model in which all modes interact with all the other modes, which ignores the situations where modes could interact in subgroups or hierarchies. Tensor network models decompose a tensor in tensor networks which expose more localized relationships between modes. Tensor networks have flexibility in modeling and compute/storage efficiency especially for high-order tensors.

TT. Tensor Train (TT) decomposition, also called Matrix Product State (MPS) in quantum physics community [27, 43], was first proposed by Ivan Oseledets in the work [101]. TT decomposes a high-order tensor into a linear sequence of tensor-times-tensor/matrix products. The contraction modes are in small rank sizes in low-rank approximation.

The variants of TT include tensor chain (TC), tensor networks with cycles: Projected Entangled Pair States (PEPS) [100], Projected Entangled Pair Operators (PEPO) [38], Honey-Comb Lattice (HCL) [40], Multi-scale Entanglement Renormalization Ansatz (MERA) [100].

The computational tensor kernels of TT are the Tensor-Scalar (Ts), Tensor-Times-Matrix (T_{TM}) and Tensor-Times-Tensor (T_{TT}) operations. Ts and T_{TM} will be described in Section 4.2 and 4.4 respectively, and T_{TT} will be one of our future work.

hTucker. Hierarchical Tucker (hTucker) decomposition, also called hierarchical tensor representation, was introduced in [27, 42–44]. hTucker recursively splits the set of tensor modes, resulting a binary tree containing a subset of modes at each node. This binary tree is called dimension tree,

and the modes from different nodes do not overlap. TT decomposition is a special case of hTucker while the dimension tree is linear and extremely unbalanced.

Variants of hTucker include the Tree Tensor Network States (TTNS) model [95], multilayer multi-configuration time-dependent Hartree method (ML-MCTDH) [150]. Sparsity has been considered by Perros et al. in the work [108].

The computational tensor kernels of hTucker are the Tensor-Scalar (Ts), Tensor-Times-Matrix (T_{TM}) and Tensor-Times-Tensor (T_{TT}) operations. Ts and T_{TM} will be described in Section 4.2 and 4.4 respectively, and T_{TT} will be one of our future work.

3.1.3 Tensor Regression. Tensor regression is an extension of classical regression model, but using tensors to represent input and covariates data. Tensor regression approximates coefficient tensor with a low-rank decomposition, thus tensor decomposition methods introduced above can be easily adopted here. Some tensor regression methods have been proposed [116, 123, 153, 157, 158, 162, 165].

3.2 Tensor Applications

Tensor methods can be used in applications to expose the inherent relationship in the observed data and to represent the data in a more compressed way. This section does not keen to give a thorough survey of tensor applications but emphasizes on showing the broad application scenarios tensor methods can be applied and useful in. Please refer to these surveys for more complete tensor applications [5, 26–28, 31, 74, 121].

3.2.1 Machine Learning. The diversity needs of machine learning algorithms have promoted the exploitation of various tensor-based decompositions, regressions, and techniques from this community. In particular, the latent variable model, where hidden factors are assumed to express structure in observed data, has been frequently expressed using CPD [55], tensor power method [5], hTucker [135], and other formats [58].

CPD, Tucker and TT decompositions have been leveraged in the context of neural networks [56, 59, 79, 96, 97, 120, 131, 156, 159], with the weight matrix of a fully-connected layer or a convolutional layer stored compressedly in a low-rank tensor, thus reducing redundancies in the network parameterization. As concerns improving theoretical aspects and understanding of deep neural networks through tensors, Cohen et al. [29] analyzed the expressive power of deep architectures by drawing analogies between shallow networks and the rank-1 CPD, as well as between deep networks and the hTucker decomposition. Novikov et al. applied TT in Google's TensorFlow [1, 96] which expresses a wide variety of algorithms as operators (graph nodes) that communicate tensor objects through the graph's edges.

Other Machine Learning applications include using TT to improve Markov Random Field (MRF) inference problem [98] and extending standard Machine Learning algorithms such as Support Vector Machines and Fisher discriminant analysis to handle tensor-based input [144]. Tensor methods involving other machine learning tasks such as feature selection and multi-way clustering will be discussed in other applications below.

3.2.2 Healthcare Analytics. The work on tensor-based healthcare data analysis has been driven by the need of improving the interpretability and the robustness of underlying methods, with the goal that healthcare professionals may eventually use consulting tools based on these methods. As a result, recent work has focused on modifying traditional tensor methods like CPD by adding constraints that better describe the underlying data and exploit domain knowledge. One particular focus is handling *sparsity*, which is particularly important when handling event-recording tensors describing healthcare data [52–54, 92, 108, 152, 165].

3.2.3 Social Network Analysis. Some studies have been done on DBLP authorship data [102] by using dynamic/static tensor analysis (include CPD, Tucker decompositions and their variants) to demonstrate clustering [76, 141], find interesting events (or anomalies) in the users' social activities [104, 105]. Jiang et al. identified patterns in human behavior through a dynamic tensor decomposition of user interactions within a microblogging service [61]. Sun et al. demonstrated a sampling-based Tucker decomposition [140], to jointly model the sender-recipient interaction and share content within business networks. The work in [10] utilizes tensors to model higher-order structures, such as cycles or feed-forward loops in a graph clustering framework.

3.2.4 Quantum Chemistry. Tensors have a long history in quantum chemistry because of the nature of high-dimensional data there [71]. Hartree–Fock (HF) is a method of approximation for the energy of a quantum many-body system and large-scale electronic structure calculations. Koppl et al. proposed sparsity using local density fitting in Hartree–Fock calculations, which heavily involves TTT and TTM operations [77]. Lewis et al. introduced a clustered low-rank tensor format to exploit element and rank sparsities [81]. Block sparsity has been utilized in coupled-cluster singles and doubles (CCSD) in the work [15, 37, 64, 91, 107]. Scaled opposite spin second order Mller-Plesset perturbation theory (SOS-MP2) method uses tensor hypercontraction (THC), approximating a electron Coulomb repulsion integrals (ERI) tensor by decomposing into lower order tensors, with sparsity [133].

3.2.5 Data Mining. Tensor decompositions have become a standard approach in brain signal analysis due to multiple heterogeneous data sources. Some recent methods have been surveyed in [17, 25]. Electroencephalogram (EEG) and fMRI data are treated as tensors and analyzed by different tensor decompositions (e.g., CPD) to study the structure of epileptic seizures [2, 3], better understand the active brain regions and their behavior [30, 78], do feature selection [16], and model neuroimaging data [94]. BrainQ is a widely available tensor dataset consisting of a sparse tensor with (subject, brain-voxel, noun) as dimensions and a matrix (noun, properties), which are measured from brain activity where individual subjects are shown nouns. Factorizing this is known as a coupled factorization [4], and Papalexakis, et al. demonstrated a scalable method using random sampling [103]. On the supervised learning setting, F. Wang et al. used fMRI data and adapted the Sparse Logistic Regression to accept tensor input that consequently avoided the loss of correlation information among different orders [149].

Personalized web search tailors the results of a search query for a particular user by utilizing the click history of this user's previous search results. Researchers constructed tensors from (user, query, webpage) information and used CPD [75] and Tucker decompositions [142] to tackle this problem.

Recommendation systems have also found tensor methods effective to resolve overloaded tags. Some approaches have been explored using CPD and Tucker decompositions and their variants on collaborative filtering [155], a tag-recommendation engine [67, 114, 143], personalized tags [39], and sparse international relationships [118].

3.2.6 Signal Processing. There has been an extensive research from the Signal Processing community, which examines theoretical aspects of tensor methods [62] such as identifiability, or improves existing decompositions [14, 122]. A tutorial addressing signal processing applications can be found in [28]. Please refer to the survey [121] for more complete applications in signal processing.

3.2.7 Other Areas. The usage of tensors and tensor decompositions as tools facilitating the extraction of useful information out of complex data is not limited to the categories mentioned above. For example, Benson, et al. used Tucker decomposition to compress scientific data obtained

by Direct Numerical Simulation (DNS) [6]. Song et al. applied CPD to forecast of the power demand and detect anomalies in smart electrical grid [134]. A variant of Tucker decomposition was used in AC optimal power flow in the work [99]. TT was used in the hierarchical uncertainty quantification to reduce the computational cost of circuit simulation [161]. Electronic design automation (EDA) problems employed CPD, Tucker, and TT decompositions to ease the suffer of the curse of dimensionality [160]. Motion control problems in the context of robotics took TT into consider for its compressed representations [41].

4 BENCHMARK WORKLOADS

This section we describe the workloads in PASTA, which includes element-wise addition/subtraction/multiplication/division, tensor-scalar, tensor-times-vector, tensor-times-matrix, and tensor-times-matrix sequence operations. We referred to the surveys [5, 26–28, 31, 74, 121] and papers [83] for these definitions.

A tensor, abstractly defined, is a function of three or more indices. In computational data analytics, one may regard a tensor as a multidimensional array, where each of its dimensions is also called a *mode* and the number of dimensions or modes is its *order*. For example, a scalar is a tensor of order 0; a vector is a tensor of order 1; and a matrix, order 2, with two modes (its rows and its columns). Notationally, we represent tensors as calligraphic capital letters, e.g., $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$; matrices by boldface capital letters, e.g., $\mathbf{U} \in \mathbb{R}^{I \times J}$; vectors by boldface lowercase letters, e.g., $\mathbf{x} \in \mathbb{R}^I$; and scalars by lowercase letters, such as x_{ijk} for the (i, j, k) element of a third-order tensor \mathcal{X} . A *slice* is a two-dimensional cross-section of a tensor, achieved by fixing all mode indices but two, e.g., $S_{::k} = \mathcal{X}(:, :, k)$ in MATLAB notation. A *fiber* is a vector extracted from a tensor along some mode, selected by fixing all indices but one, e.g., $\mathbf{f}_{:jk} = \mathcal{X}(:, j, k)$.

A tensor can be reshaped to a matrix, which is called matricization. For a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \dots \times I_n \times \dots \times I_N}$, its matricized tensor along with mode- n is $\mathbf{X}_{(n)} \in \mathbb{R}^{I_1 \dots I_{n-1} I_{n+1} \dots I_N \times I_n}$. A matrix can be also reshaped to a tensor by splitting one mode into two or more.

4.1 Tensor Element-Wise Operations

Tensor element-wise (TEW) operations include addition, subtraction, multiplication, and division operations, which are applied to every corresponding pair of elements from two tensor objects if they have the same order and shape (dimension sizes). For example, element-wise tensor addition of $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ is $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$, where

$$z_{i_1 \dots i_N} = x_{i_1 \dots i_N} + y_{i_1 \dots i_N}. \quad (1)$$

Similarly for element-wise tensor subtraction $\mathcal{Z} = \mathcal{X} - \mathcal{Y}$, multiplication $\mathcal{Z} = \mathcal{X} * \mathcal{Y}$, and division $\mathcal{Z} = \mathcal{X} / \mathcal{Y}$. When the two input tensors have exactly the same non-zero distribution, element-wise operations can be easily implemented by iterating all non-zeros of the two sparse tensors and doing the corresponding operation for each element. The tricky cases are when the non-zero patterns of tensors \mathcal{X} and \mathcal{Y} are different and even worse they could be in different shapes. For these two cases, we cannot easily predict the output tensor \mathcal{Z} 's storage space before computation. These two cases we use dynamic vectors and an optimization strategy for parallel algorithms.

4.2 Tensor-Scalar Operations

A Tensor-Scalar (Ts) operation is the addition (TsA) /subtraction (TsS) /multiplication (TSM) /division (TSD) of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_N}$ with a scalar $s \in \mathbb{R}$ for every non-zero entry. It is denoted by $\mathcal{Y} = \mathcal{X} \times s$. For example, the TSM operation is defined as

$$y_{i_1 \dots i_{n-1} i_{n+1} \dots i_N} = s \times x_{i_1 \dots i_{n-1} i_n i_{n+1} \dots i_N}. \quad (2)$$

Since $\mathcal{Y} = \mathcal{X} \times s$ is the same with $\mathcal{Y} = \mathcal{X}/s^{-1}$ and $\mathcal{Y} = \mathcal{X} + s$ is the same with $\mathcal{Y} = \mathcal{X} - (-s)$, so implementing TSA and TSM is enough.

4.3 Tensor-Times-Vector Operation

The Tensor-Times-Vector (TTV) in mode n is the multiplication of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \dots \times I_n \times \dots \times I_N}$ with a vector $\mathbf{v} \in \mathbb{R}^{I_n}$, along mode n , and is denoted by $\mathcal{Y} = \mathcal{X} \times_n \mathbf{v}$. This results in a $I_1 \times \dots \times I_{n-1} \times I_{n+1} \times \dots \times I_N$ tensor which has one less dimension. Its operation is defined as

$$y_{i_1 \dots i_{n-1} i_{n+1} \dots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 \dots i_{n-1} i_n i_{n+1} \dots i_N} v_{i_n}. \quad (3)$$

4.4 Tensor-Times-Matrix Operation

The Tensor-Times-Matrix (TTM) in mode n , also known as the n -mode product, is the multiplication of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \dots \times I_n \times \dots \times I_N}$ with a matrix $\mathbf{U} \in \mathbb{R}^{I_n \times R}$, along mode n , and is denoted by $\mathcal{Y} = \mathcal{X} \times_n \mathbf{U}$.¹ This results in a $I_1 \times \dots \times I_{n-1} \times R \times I_{n+1} \times \dots \times I_N$ tensor, and its operation is defined as

$$y_{i_1 \dots i_{n-1} r i_{n+1} \dots i_N} = \sum_{i_n=1}^{I_n} x_{i_1 \dots i_{n-1} i_n i_{n+1} \dots i_N} u_{i_n r}. \quad (4)$$

TTM is a special case of tensor contraction. We consider TTM specifically because of its more common usage in tensor decompositions for data analysis, such as the Tucker decomposition. Also, note that R is typically much smaller than I_n in such decompositions, and typically $R < 100$.

TTM is also equivalent to a matrix-matrix multiplication in the following form:

$$\mathcal{Y} = \mathcal{X} \times_n \mathbf{U} \quad \Leftrightarrow \quad \mathbf{Y}_{(n)} = \mathbf{U} \mathbf{X}_{(n)}. \quad (5)$$

Therefore, one feasible way to implement an TTM is to first matricize the tensor, then use an optimized matrix-matrix multiplication to compute the matricized output \mathcal{Y} , and, finally, tensorize to obtain \mathcal{Y} . However it has the tensor-matrix transformation as the extra overhead and does not work well for sparse tensors.

4.5 Kronecker and Khatri-Rao Products

Kronecker and Khatri-Rao products are both matrix products. The *Kronecker product* generalizes the outer product for matrices. Given $\mathbf{U} \in \mathbb{R}^{I \times J}$ and $\mathbf{V} \in \mathbb{R}^{K \times L}$, the Kronecker product $\mathbf{U} \otimes \mathbf{V} \in \mathbb{R}^{IK \times JL}$ is

$$\mathbf{U} \otimes \mathbf{V} = \begin{bmatrix} u_{11}\mathbf{V} & u_{12}\mathbf{V} & \dots & u_{1J}\mathbf{V} \\ u_{21}\mathbf{V} & u_{22}\mathbf{V} & \dots & u_{2J}\mathbf{V} \\ \vdots & \vdots & \ddots & \vdots \\ u_{I1}\mathbf{V} & u_{I2}\mathbf{V} & \dots & u_{IJ}\mathbf{V} \end{bmatrix} \quad (6)$$

The *Khatri-Rao product* is a “matching column-wise” Kronecker product between two matrices with the same number of columns. Given matrices $\mathbf{A} \in \mathbb{R}^{I \times R}$ and $\mathbf{B} \in \mathbb{R}^{J \times R}$, their Khatri-Rao product is denoted by $\mathbf{A} \odot \mathbf{B} \in \mathbb{R}^{(IJ) \times R}$,

$$\mathbf{A} \odot \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1, \mathbf{a}_2 \otimes \mathbf{b}_2, \dots, \mathbf{a}_R \otimes \mathbf{b}_R], \quad (7)$$

where \mathbf{a}_r and \mathbf{b}_r , $r = 1, \dots, R$, are columns of \mathbf{A} and \mathbf{B} .

¹Our convention for the dimensions of \mathbf{U} differs from that of Kolda and Bader’s definition [74]. In particular, we transpose the matrix modes \mathbf{U} , which leads to a more efficient TTM under the row-major storage convention of the C language.

Kronecker and Khatri-Rao products appear frequently in tensor decompositions that are formulated as matrix operations. However, such formulations typically also require redundant computation or extra storage to hold matrix operands, so in practice these operations tend to be not implemented directly but rather integrated into tensor operations.

4.6 Tensor-Times-Matrix Sequence Operation

There are two types of tensor-times-matrix sequence operations, TTM chain and MTTKRP. TTM chain is a sequence of TTM operations with one's output as the next one's input. An alternative way to think TTM chain is a matricized tensor times the Kronecker product of matrices. MTTKRP, matricized tensor times Khatri-Rao product, is a matricized tensor times the Khatri-Rao product of matrices. For an N^{th} -order tensor \mathcal{X} and given matrices $\mathbf{U}^{(1)}, \dots, \mathbf{U}^{(N)}$, the mode- n MTTKRP is

$$\tilde{\mathbf{U}}^{(n)} = \mathbf{X}_{(n)} \left(\odot_{i=1, \dots, N}^{i \neq n} \mathbf{U}_i \right) = \mathbf{X}_{(n)} \left(\mathbf{U}^{(N)} \odot \dots \odot \mathbf{U}^{(n+1)} \odot \mathbf{U}^{(n-1)} \odot \dots \odot \mathbf{U}^{(1)} \right), \quad (8)$$

where $\mathbf{X}_{(n)}$ is the mode- n matricization of tensor \mathcal{X} , \odot is the Khatri-Rao product.

4.7 Others

We also provide the transformation between tensors and matrices and some sorting algorithms for sparse tensors.

5 DATA STRUCTURES, ALGORITHMS, AND IMPLEMENTATIONS

5.1 Data Structures

Since COO [74] is the simplest and arguably de facto standard way to store a sparse tensor, and it is mode generic, we only support COO format in this work. Other state-of-the-art formats will be included as our future work. We use inds and val to represent the indices and values of the non-zeros of a sparse tensor respectively. val is a size- M array of floating-point numbers, inds is a size- M array of integer tuples. Figure 1 shows a $4 \times 4 \times 3$ sparse tensor in COO format. The indices of each mode are represented as i, j , and k . Observe that some indices in inds repeat, for example, entries $(1, 0, 0)$ and $(1, 0, 2)$ have the same i and j indices. This redundancy suggests some compression of this indexing metadata should be possible, as proposed in some work [89, 130].

i	j	k	val
0	0	0	1
0	1	0	2
1	0	0	3
1	0	2	4
2	1	0	5
2	2	2	6
3	0	1	7
3	3	2	8

Fig. 1. COO format of an example $4 \times 4 \times 3$ tensor.

5.2 Algorithms

This section describes the sequential algorithms for the workloads in Section 4. All algorithms directly operates on the input sparse tensor(s) without explicit tensor-matrix transformation.

Algorithm 1 Sequential COO-Tew-eq-Addition algorithm for tensors in the same order and shape.**Input:** A third-order sparse tensor $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{I \times J \times K}$ with M non-zeros;**Output:** Sparse tensor $\mathcal{Z} \in \mathbb{R}^{I \times J \times K}$;

-
- ```

1: Allocate \mathcal{Z} space with M non-zeros; ▷ $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$
2: for $m = 1, \dots, M$ do ▷ Pre-allocation space.
3: $\text{inds}_z^1(m) = \text{inds}_x^1(m), \text{inds}_z^2(m) = \text{inds}_x^2(m), \text{inds}_z^3(m) = \text{inds}_x^3(m)$;
4: $\text{val}_z(m) = \text{val}_x(m) + \text{val}_y(m)$;
5: return \mathcal{Z} ;

```
- 

5.2.1 *TEW*. As mentioned in Section 4.1, TEW operation has two cases: one is between two tensors in exactly the same shape and non-zero distribution; the other only requires the two tensors are in the same tensor order.

For the first case, we show TEW addition as an example in Algorithm 1. The output tensor has the same shape and non-zero distribution with the two input tensors, thus it can be pre-allocated. Then the calculation simply does addition by looping all non-zeros.

**Algorithm 2** Sequential COO-Tew-Addition algorithm for general tensors.**Input:** A third-order sparse tensor  $\mathcal{X} \in \mathbb{R}^{I_1 \times J_1 \times K_1}$  with  $M_1$  non-zeros,  $\mathcal{Y} \in \mathbb{R}^{I_2 \times J_2 \times K_2}$  with  $M_2$  non-zeros;**Output:** Sparse tensor  $\mathcal{Z} \in \mathbb{R}^{I_3 \times J_3 \times K_3}$ ;

- 
- ```

1:  $I_3 = \max\{I_1, I_2\}, J_3 = \max\{J_1, J_2\}, K_3 = \max\{K_1, K_2\}$  ▷  $\mathcal{Z} = \mathcal{X} + \mathcal{Y}$ 
2: Sort  $\mathcal{X}$  and  $\mathcal{Y}$  in the same dimension order. ▷ Unify the tensor shape.
3:  $m_1 = 1, m_2 = 1$ 
4: while  $m_1 < M_1$  and  $m_2 < M_2$  do
5:   if  $\text{inds}_x == \text{inds}_y$  then
6:     Append( $\text{inds}_z^1, \text{inds}_x^1(m_1)$ ); Append( $\text{inds}_z^2, \text{inds}_x^2(m_1)$ ); Append( $\text{inds}_z^3, \text{inds}_x^3(m_1)$ );
7:     Append( $\text{val}_z, \text{val}_x(m_1) + \text{val}_y(m_2)$ );
8:   if  $\text{inds}_x > \text{inds}_y$  then
9:     Append( $\text{inds}_z^1, \text{inds}_y^1(m_1)$ ); Append( $\text{inds}_z^2, \text{inds}_y^2(m_1)$ ); Append( $\text{inds}_z^3, \text{inds}_y^3(m_1)$ );
10:    Append( $\text{val}_z, \text{val}_y(m_2)$ );
11:   if  $\text{inds}_x < \text{inds}_y$  then
12:     Append( $\text{inds}_z^1, \text{inds}_x^1(m_1)$ ); Append( $\text{inds}_z^2, \text{inds}_x^2(m_1)$ ); Append( $\text{inds}_z^3, \text{inds}_x^3(m_1)$ );
13:     Append( $\text{val}_z, \text{val}_x(m_1)$ );
14:   if  $m_1 < M_1$  then
15:     Append( $\text{inds}_z^1, \text{inds}_x^1(m_1, :)$ ); Append( $\text{inds}_z^2, \text{inds}_x^2(m_1, :)$ ); Append( $\text{inds}_z^3, \text{inds}_x^3(m_1, :)$ );
16:     Append( $\text{val}_z, \text{val}_x(m_1, :)$ );
17:   if  $m_2 < M_2$  then
18:     Append( $\text{inds}_z^1, \text{inds}_x^1(m_2, :)$ ); Append( $\text{inds}_z^2, \text{inds}_x^2(m_2, :)$ ); Append( $\text{inds}_z^3, \text{inds}_x^3(m_2, :)$ );
19:     Append( $\text{val}_z, \text{val}_x(m_2, :)$ );
20: return  $\mathcal{Z}$ ;

```
-

For the second case, its algorithm is shown in Algorithm 2. The output tensor size is set by the maximum dimension size of the two input tensors. Since we do not know the number of the output non-zeros, we cannot pre-allocate the space of the output tensor \mathcal{Z} but using dynamic allocation to append non-zeros. First, we need to sort tensors \mathcal{X} and \mathcal{Y} in the order of mode $1 > 2 > 3$, then compare the indices in lexicographical order for each non-zero pair-to-pair, e.g.,

indices $(2, 1, 1) > (1, 1, 2) > (1, 1, 1)$. If two indices are the equal, then we append the indices and the sum of the two non-zero values to the output \mathcal{Z} . Otherwise, we append the smaller indices and its corresponding value to \mathcal{Z} . Only if we run out of non-zeros in either \mathcal{X} or \mathcal{Y} , we append the rest indices and values of the other one to \mathcal{Z} .

Algorithm 3 Sequential COO-Tsm algorithm.

Input: A third-order sparse tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ with M non-zeros;

Output: Output sparse tensor $\mathcal{Y} \in \mathbb{R}^{I \times J \times K}$;

- 1: Allocate \mathcal{Y} space with M non-zeros; ▷ $\mathcal{Y} = \mathcal{X} \times s$
 - 2: **for** $m = 1, \dots, M$ **do** ▷ Pre-allocation space.
 - 3: $\text{inds}_y^1(m) = \text{inds}_x^1(m), \text{inds}_y^2(m) = \text{inds}_x^2(m), \text{inds}_y^3(m) = \text{inds}_x^3(m)$;
 - 4: $\text{val}_y(m) = s \times \text{val}_x(m)$;
 - 5: **return** \mathcal{Y} ;
-

5.2.2 *Ts*. Ts algorithm is simple. The output \mathcal{Y} can be pre-allocated and computed by looping all non-zeros. Algorithm 3 shows the Tsm algorithm.

Algorithm 4 Sequential COO-Ttv algorithm.

Input: A third-order sparse tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, dense vector $\mathbf{V} \in \mathbb{R}^K$, mode $n = 3$;

Output: Sparse tensor $\mathcal{Y} \in \mathbb{R}^{I \times J}$;

- 1: Pre-process to obtain M_F : the number of mode- n fibers of \mathcal{X} and f_{ptr} : the beginnings of each \mathcal{X} mode- n fiber, sized M_F .
 - 2: Allocate \mathcal{Y} space with M_F non-zeros; ▷ $\mathcal{Y} = \mathcal{X} \times_n \mathbf{v}$
 - 3: **for** $f = 1, \dots, M_F$ **do** ▷ Pre-allocation space.
 - 4: $\text{inds}_y^1(f) = \text{inds}_x^1(f_{\text{ptr}}(f)), \text{inds}_y^2(f) = \text{inds}_x^2(f_{\text{ptr}}(f))$
 - 5: **for** $m = f_{\text{ptr}}(f), \dots, f_{\text{ptr}}(f + 1) - 1$ **do**
 - 6: $k = \text{inds}_x^3(m)$
 - 7: $\text{val}_y(f) += \text{val}_x(m) \times u(k)$
 - 8: **Return** \mathcal{Y} ;
-

5.2.3 *Ttv*. Ttv algorithm in mode- n is shown in Algorithm 4. It first pre-compute the number of fibers M_F of input tensor \mathcal{X} and the beginning positions of each fiber. Then we can pre-allocate the output tensor \mathcal{Y} with M_F , because this product does not influence the non-zero layout for I and J modes. The algorithm loops all the fibers of \mathcal{X} , and a reduction happens for all non-zeros in each fiber.

5.2.4 *TTM*. TTM algorithm is illustrated in Algorithm 5. Similarly to Ttv algorithm, we obtain the number of fibers M_F and the beginning positions of each fiber then $M_F \times R$ space are allocated for the output tensor \mathcal{Y} . The algorithm loops all the M_F fibers and does a reduction between sized- R vectors. This TTM algorithm directly operates on the input sparse tensor by avoiding tensor transformation. The explanation of Algorithm 5 can be found in the work [85, 90].

5.2.5 *MTTKRP*. MTTKRP algorithm is shown in Algorithm 6, the output matrix of which is initialized before and only needs to be updated. This algorithm loops all non-zeros of the tensor \mathcal{X} and times the corresponding two matrix vectors, to update the designated output matrix vector. Readers can refer more details of this algorithm in [7].

Algorithm 5 Sequential COO-TTM algorithm [85].**Input:** A sparse tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, a dense matrix $U \in \mathbb{R}^{K \times R}$, and an integer $n = 3$;**Output:** Sparse tensor $\mathcal{Y} \in \mathbb{R}^{I \times J \times R}$;

▷ $\mathcal{Y} = \mathcal{X} \times_n U$

```

1: Pre-process to obtain  $M_F$ : the number of mode- $n$  fibers of  $\mathcal{X}$  and  $f_{\text{ptr}}$ : the beginnings of each  $\mathcal{X}$  mode- $n$ 
   fiber, size  $M_F$ .
2: Allocate  $\mathcal{Y}$  space with  $M_F \times R$  non-zeros;                                ▷ Pre-allocation space.
3: for  $f = 1, \dots, M_F$  do
4:    $i = \text{inds}_X^1(f_{\text{ptr}}(f)), j = \text{inds}_X^2(f_{\text{ptr}}(f))$ 
5:   for  $r = 1, \dots, R$  do
6:      $\text{inds}_Y^1(f \times R + r) = i, \text{inds}_Y^2(f \times R + r) = j, \text{inds}_Y^3(f \times R + r) = r$ 
7:   for  $m = f_{\text{ptr}}(f), \dots, f_{\text{ptr}}(f+1) - 1$  do
8:      $k = \text{inds}_X^3(m)$ 
9:      $value = \text{val}_X(m)$ 
10:    for  $r = 1, \dots, R$  do
11:       $\text{val}_Y(f \times R + r) += value \times u(k \times R + r)$ 
12: Return  $\mathcal{Y}$ ;
```

Algorithm 6 Sequential COO-MTTKRP algorithm ([7]).**Input:** A third-order sparse tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, dense matrices $B \in \mathbb{R}^{J \times R}, C \in \mathbb{R}^{K \times R}$;**Output:** Updated dense matrix $\tilde{A} \in \mathbb{R}^{I \times R}$;

▷ $\tilde{A} \leftarrow \mathcal{X}_{(1)}(C \odot B)$

```

1: for  $m = 1, \dots, M$  do
2:    $i = \text{inds}_X^1(m), j = \text{inds}_X^2(m), k = \text{inds}_X^3(m)$ ;
3:    $value = \text{val}(m)$ 
4:   for  $r = 1, \dots, R$  do
5:      $\tilde{A}(i \times R + r) += value \times C(k \times R + r) \times B(j \times R + r)$ 
6: return  $\tilde{A}$ ;
```

Table 2. The analysis of data storage and their algorithms for third-order cubical tensors ($\mathcal{X} \in \mathbb{R}^{I \times I \times I}$). We consider all input tensors with M nonzero entries and M_F fibers, $I \ll M_F \ll M$. The indices use 32 bits, and values are single-precision floating-point numbers with 32 bits.

Workloads	Storage (Bytes)	Work (Flops)	Memory Access (Bytes)	Arithmetic Intensity (AI)
TEW	$48M$	M	$36M$	$1/36$
TS	$32M$	M	$32M$	$1/32$
TTV	$(16M + 12M_F)$	$2M$	$(12M + 20M_F)$	$\sim 1/6$
TTM	$(16M + 16M_F R + 4IR)$	$2MR$	$4MR + 8M + 12M_F R + 8M_F$	$\sim 1/2$
MTTKRP	$(16M + 12IR)$	$3MR$	$12MR + 16M$	$\sim 1/4$

According to the above algorithms, we compute the storage, the number of floating-point operations (Flops), the amount of memory access in bytes, and the arithmetic intensity (the ratio of #Flops/#Bytes) in Table 2. For simplicity, we use a cubical third-order sparse tensor $\mathcal{X} \in \mathbb{R}^{I \times I \times I}$ with M non-zeros and M_F fibers as an example. Because of the irregular access pattern of sparse tensors, the memory access does not consider the cache effect. All workloads have arithmetic intensity less than 1, thus it is hard to easily achieve good performance on common architectures.

While MTTKRP has the most Flops and memory access, its arithmetic intensity is smaller than TTM, which is $\sim 1/2$. TEW and Ts have the smallest arithmetic intensity and the largest storage due to the output tensor. Despite of different algorithm behavior, these algorithms are generally considered memory intensive, which demonstrates the emphasis of our PASTA.

5.3 Multicore Implementations

Some workloads are easy to parallelize. We parallelize the loop of all non-zeros in TEW-eq (Algorithm 1) and Ts (Algorithm 3). For TTV (Algorithm 4) and TTM (Algorithm 5), the loop of fibers is parallelized because each fiber computation is independent.

TEW (Algorithm 2) is difficult to be parallelized because of its dynamic append operations and no pre-allocation available. We partition the two tensors in such a way that there is no overlap between their indices, then we run TEW algorithm locally for a sub-tensor in each thread and append the results to a local output buffer. The partitioning first split one of the two tensors (say \mathcal{X}) by slices and meanwhile tend to evenly distribute its non-zeros. This makes sure that all non-zeros of a slice cannot be split into two partitions. Then the partitioning of the other tensor (say \mathcal{Y}) is according to this slice partitioning strategy. In this case, we assure every partition does not overlap with each other, thus they can independently computed in parallel.

We parallelize the loop of all non-zeros of MTTKRP (Algorithm 6) as well, but Line 4 may have data race by writing into the same location of $\tilde{\mathbf{A}}$. We implemented two solutions: 1) Use atomics to protect the correctness, but the performance suffers much; 2) Employ privatization approach to allocate a thread-local buffer. The data is first written to this buffer by each thread privately, then a global reduction for the buffers is used to get the final results. In this case, we can generally get better performance than using atomics.

For these parallel implementations, we have not considered the NUMA effect, which will be another piece of our future work.

6 DATASET

PASTA now only considers real-world data as input. The sparse tensors derived from real-world applications, that appear in Table 3, ordered by decreasing non-zero density separately for third- and fourth-order tensors. Most of these tensors are included in The Formidable Repository of Open Sparse Tensors and Tools (FROSTT) dataset (Refer to the details in [125]). The darpa (source IP-destination IP-time triples), fb-m, and fb-s (short for “freebase-music” and “freebase-sampled”, entity-entity-relation triples) are from the dataset of HaTen2 [60], and choa is built from electronic health records (EHRs) of pediatric patients at Children’s Healthcare of Atlanta (CHOA) [109].

7 EXPERIMENTS

We tested these schemes experimentally on a Linux-based Intel Xeon E5-2698 v3 multicore server platform with 32 physical cores distributed on two sockets, each with 2.3 GHz frequency. The processor microarchitecture is Haswell, having 32 KiB L1 data cache and 128 GiB memory. The code artifact is written in the C language using OpenMP parallelization, and was compiled using icc 18.0.1. All experiments use 32 threads for parallel code except being pointed out otherwise. The execution time are all averaged by five runs. For TTM and MTTKRP, we set the rank $R = 16$.

We demonstrate the sequential and multicore parallel performance for every workload on the dataset (Table 3).

Table 3. Description of sparse tensors.

Tensors	Order	Dimensions	#Non-zeros	Density
vast	3	$165K \times 11K \times 2$	26M	6.9×10^{-3}
nell2	3	$12K \times 9K \times 29K$	77M	2.4×10^{-5}
choa	3	$712K \times 10K \times 767$	27M	5.0×10^{-6}
darpa	3	$22K \times 22K \times 24M$	28M	2.4×10^{-9}
fb-m	3	$23M \times 23M \times 166$	100M	1.1×10^{-9}
fb-s	3	$39M \times 39M \times 532$	140M	1.7×10^{-10}
deli	3	$533K \times 17M \times 2.5M$	140M	6.1×10^{-12}
nell1	3	$2.9M \times 2.1M \times 25M$	144M	9.1×10^{-13}
crime	4	$6K \times 24 \times 77 \times 32$	5M	1.5×10^{-2}
nips	4	$2K \times 3K \times 14K \times 17$	3M	1.8×10^{-6}
enron	4	$6K \times 6K \times 244K \times 1K$	54M	5.5×10^{-9}
flickr4d	4	$320K \times 28M \times 1.6M \times 731$	113M	1.1×10^{-14}
deli4d	4	$533K \times 17M \times 2.5M \times 1K$	140M	4.3×10^{-15}

7.1 TEW

Figure 2 and 3 show the execution time of the two cases of TEW addition (Algorithm 1 and 2): in the same non-zero pattern and only in the same tensor order, on all third- and fourth-order tensors. We use the same tensor for the two input for TEW-eq and TEW to better show the algorithm effect. We observe for both cases, parallel TEW outperforms sequential TEW. However, the speedup of TEW-eq is $3.64 - 5.18\times$, while the speedup of TEW is much smaller, which is $1.13 - 1.70\times$. This is because: 1) the parallel strategy of TEW could have a lot more load imbalance than TEW-eq's even non-zero parallelization; 2) some tensors cannot fully use all 32 threads due to the slice partitioning (a heavy slice cannot be further partitioned in Algorithm 2). Besides, due to the dynamic append operation, the sequential TEW is tens of times slower than sequential TEW-eq. From our experiments, TEW subtraction, multiplication, and division behave very similar to TEW addition in execution time.

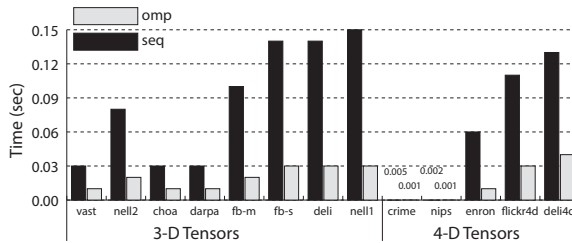


Fig. 2. TEW-eq-addition for sparse tensors in the same shape and non-zero pattern.

7.2 Ts

Figure 4 plots the sequential and parallel execution time of Tsm. Parallel Tsm achieves $2.17 - 5.92\times$ speedup over sequential Tsm, this is comparable to TEW-eq in Figure 2. The sequential Tsm executes faster than the sequential TEW, which verifies the analysis in Table 2 and that these two algorithms are memory-bound. (Because they have the same #Flops, compute-bound algorithms should have similar execution time.) From the experiments, the execution times of sequential and parallel Ts are very close to Tsm.

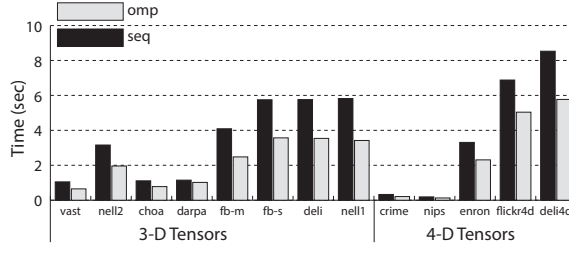


Fig. 3. Tew-addition for sparse tensors in the same order.

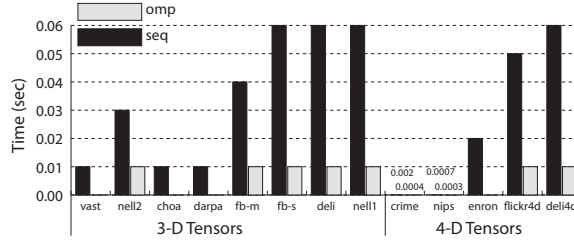


Fig. 4. Tsm execution time.

7.3 Ttv

We illustrate sequential and parallel TTV time in Figure 5. Parallel TTV outperforms sequential case by 5.21 – 12.45 \times , this is much higher than the speedup of TEW-eq, TEW, and Tsm. This behavior again matches the analysis in Table 2 that TTV has higher arithmetic intensity. Since higher arithmetic intensity potentially generates less memory contention, thus multicore parallelization could benefit more.

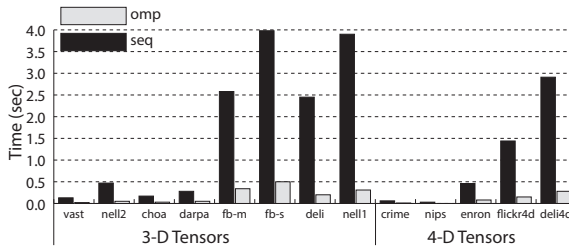


Fig. 5. Ttv: the sum of execution time of all the modes.

7.4 Ttm

Figure 6 shows the sequential and parallel execution time of TTM. The speedup of parallel TTM over sequential case is 4.09 – 15.67 \times which is comparable with TTV's. This also verifies the analysis that TTM has the highest arithmetic intensity. Sequential TTM is 4.91 – 11.11 \times slower than sequential TTV, that shows the different behavior of timing a dense vector versus a dense matrix.

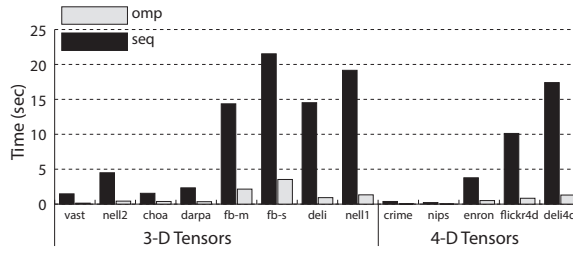


Fig. 6. TTM: the sum of execution time of all the modes.

7.5 MTTKRP

We use privatization technique for parallel MTTKRP, because it performs better than atomics technique on most of tensors. The execution time of sequential and parallel MTTKRP is shown in Figure 7, where the parallel case gains $0.77 - 9.49\times$ speedup. For tensor darpa, the only case parallel MTTKRP is slower than sequential one because of its large thread-local buffer which consumes a large portion of time to do reduction. The atomics parallel approach could be better in this case, 7.93 versus 7.32 (sequential MTTKRP), but there is still not speedup for this tensor. MTTKRP obtains smaller speedup than TTM and TTV mainly because data race exists in the output. Even we use privatization technique to avoid the data race, the extra reduction still take nontrivial amount of time.

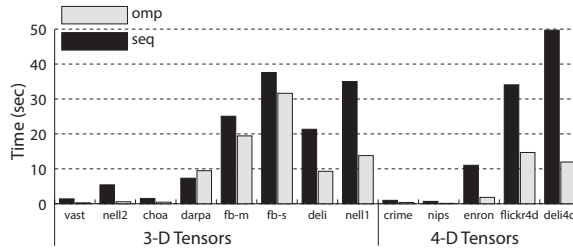


Fig. 7. MTTKRP: the sum of execution time of all the modes.

From our experiments and analysis above, these relatively simple workloads can well reflect some architecture characteristics. This can help architecture designers and application users to evaluate computer systems.

8 CONCLUSION

This work presents a sparse tensor algorithm benchmark suite (PASTA) for single-core and multi-core CPUs, which is the first sparse tensor benchmark to the best of our knowledge. PASTA consists of TEW, Ts, TTV, TTM, MTTKRP workloads to represent sparse tensor algorithms from different tensor methods in a various application scenarios. Besides, these workloads can reflect computer architecture features differently from our analysis.

As a benchmark suite, PASTA already processes good properties such as application and machine diversity, state-of-the-art data structures, algorithms, and optimization techniques included, compatibility for research support, and real-world data set. Some future work should be done to make PASTA more complete and robust: 1) more computer systems support, such as GPUs, FPGAs, and distributed systems; 2) more workloads especially tensor-times-tensor product (TTT); 3) more

state-of-the-art sparse tensor formats, e.g., hierarchical COO (HiCOO) and compressed sparse fiber (CSF) format; 4) synthetic data generation for more precise machine performance measurement.

ACKNOWLEDGMENTS

This research was partially funded by the US Department of Energy, Office for Advanced Scientific Computing (ASCR) under Award No. 66150: "CENATE: The Center for Advanced Technology Evaluation". Pacific Northwest National Laboratory (PNNL) is a multiprogram national laboratory operated for DOE by Battelle Memorial Institute under Contract DE-AC05-76RL01830.

REFERENCES

- [1] Martin Abadi et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.
- [2] Evrim Acar, Canan Aykut-Bingol, Haluk Bingol, Rasmus Bro, and Bülent Yener. 2007. Multiway Analysis of Epilepsy Tensors. *Bioinformatics* 23, 13 (July 2007), i10–i18. <https://doi.org/10.1093/bioinformatics/btm210>
- [3] Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. 2011. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems* 106, 1 (2011), 41–56.
- [4] Evrim Acar, Tamara G. Kolda, and Daniel M. Dunlavy. 2011. All-at-once Optimization for Coupled Matrix and Tensor Factorizations.
- [5] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. 2014. Tensor Decompositions for Learning Latent Variable Models. *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 2773–2832.
- [6] W. Austin, G. Ballard, and T. G. Kolda. 2016. Parallel Tensor Compression for Large-Scale Scientific Data. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 912–922. <https://doi.org/10.1109/IPDPS.2016.67>
- [7] Brett W. Bader and Tamara G. Kolda. 2007. Efficient MATLAB computations with sparse and factored tensors. *SIAM Journal on Scientific Computing* 30, 1 (December 2007), 205–231. <https://doi.org/10.1137/060676489>
- [8] Brett W. Bader, Tamara G. Kolda, et al. 2017. MATLAB Tensor Toolbox (Version 3.0-dev). Available online. <https://www.tensortoolbox.org>
- [9] Casey Battaglini, Grey Ballard, and Tamara G. Kolda. [n. d.]. A Practical Randomized CP Tensor Decomposition. *SIAM J. Matrix Anal. Appl.* 39, 2 ([n. d.]), 876–901.
- [10] Austin R Benson, David F Gleich, and Jure Leskovec. 2015. Tensor Spectral Clustering for Partitioning Higher-order Network Structures. *arXiv preprint arXiv:1502.05058* (2015).
- [11] Alex Beutel, Abhimanu Kumar, Evangelos Papalexakis, Partha Pratim Talukdar, Christos Faloutsos, and Eric P Xing. 2013. FLEXIFACT: Scalable Flexible Factorization of Coupled Tensors on Hadoop. In *NIPS 2013 Big Learning Workshop*.
- [12] Christian Bienia, Sanjeev Kumar, Jaswinder Pal Singh, and Kai Li. 2008. The PARSEC benchmark suite: Characterization and architectural implications. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*. ACM, 72–81.
- [13] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.* 3, 1 (Jan. 2011), 1–122. <https://doi.org/10.1561/22000000016>
- [14] R. Bro, N. D. Sidiropoulos, and G. B. Giannakis. 1999. A fast least squares algorithm for separating trilinear mixtures. In *Independent Component Analysis*.
- [15] Justus A. Calvin and Edward F. Valeev. 2016. TiledArray: A massively-parallel, block-sparse tensor framework (Version v0.6.0). Available from <https://github.com/valeevgroup/tiledarray>.
- [16] Bokai Cao, Lifang He, Xiangnan Kong, Philip S. Yu, Zhifeng Hao, and Ann B. Ragin. 2014. Tensor-Based Multi-view Feature Selection with Applications to Brain Diseases. In *Data Mining (ICDM), 2014 IEEE International Conference on*. 40–49. <https://doi.org/10.1109/ICDM.2014.26>
- [17] Bokai Cao, Xiangnan Kong, and Philip S. Yu. 2015. A review of heterogeneous data mining for brain disorders. *CoRR* abs/1508.01023 (2015). <http://arxiv.org/abs/1508.01023>
- [18] J. Douglas Carroll and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* 35, 3 (01 Sep 1970), 283–319. <https://doi.org/10.1007/BF02310791>
- [19] J. D. Carroll, S. Pruzansky, and J. B. Kruskal. 1980. CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika* 45 (1980), 3–24.
- [20] S. Che, M. Boyer, J. Meng, D. Tarjan, J. W. Sheaffer, S. Lee, and K. Skadron. 2009. Rodinia: A benchmark suite for heterogeneous computing. In *2009 IEEE International Symposium on Workload Characterization (IISWC)*. 44–54. <https://doi.org/10.1109/IISWC.2009.5306797>

- [21] Dehua Cheng, Richard Peng, Yan Liu, and Ioakeim Perros. 2016. SPALS: Fast Alternating Least Squares via Implicit Leverage Scores Sampling. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 721–729. <http://papers.nips.cc/paper/6436-spals-fast-alternating-least-squares-via-implicit-leverage-scores-sampling.pdf>
- [22] Eric C Chi and Tamara G Kolda. 2012. On tensors, sparsity, and nonnegative factorizations. *SIAM J. Matrix Anal. Appl.* 33, 4 (2012), 1272–1299.
- [23] Jee Choi, Xing Liu, Shaden Smith, and Tyler Simon. 2018. Blocking Optimization Techniques for Sparse Tensor Computation. 568–577. <https://doi.org/10.1109/IPDPS.2018.00066>
- [24] Joon Hee Choi and S. Vishwanathan. 2014. DFacTo: Distributed Factorization of Tensors. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger (Eds.). Curran Associates, Inc., 1296–1304.
- [25] Andrzej Cichocki. 2013. Tensor Decompositions: A New Concept in Brain Data Analysis? *arXiv preprint arXiv:1305.0395* (2013).
- [26] Andrzej Cichocki. 2014. Era of Big Data Processing: A New Approach via Tensor Networks and Tensor Decompositions. *CoRR abs/1403.2048* (2014).
- [27] A. Cichocki, N. Lee, I. V. Oseledets, A. Phan, Q. Zhao, and D. Mandic. 2016. Low-Rank Tensor Networks for Dimensionality Reduction and Large-Scale Optimization Problems: Perspectives and Challenges PART 1. *ArXiv e-prints* (Sept. 2016). [arXiv:cs.NA/1609.00893](http://arxiv.org/abs/1609.00893)
- [28] A. Cichocki, D. Mandic, L. De Lathauwer, Guoxu Zhou, Qibin Zhao, C. Caiafa, and H.A. Phan. 2015. Tensor Decompositions for Signal Processing Applications: From two-way to multiway component analysis. *Signal Processing Magazine, IEEE* 32, 2 (March 2015), 145–163. <https://doi.org/10.1109/MSP.2013.2297439>
- [29] Nadav Cohen, Or Sharir, and Amnon Shashua. 2015. On the Expressive Power of Deep Learning: A Tensor Analysis. *CoRR abs/1509.05009* (2015). <http://arxiv.org/abs/1509.05009>
- [30] Ian Davidson, Sean Gilpin, Owen Carmichael, and Peter Walker. 2013. Network Discovery via Constrained Tensor Analysis of fMRI Data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, New York, NY, USA, 194–202. <https://doi.org/10.1145/2487575.2487619>
- [31] Lieven De Lathauwer. 2008. Decompositions of a Higher-Order Tensor in Block Terms—Part I: Lemmas for Partitioned Matrices. *SIAM J. Matrix Anal. Appl.* 30, 3 (2008), 1022–1032. <https://doi.org/10.1137/060661685> [arXiv:http://dx.doi.org/10.1137/060661685](http://dx.doi.org/10.1137/060661685)
- [32] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. 2000. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* 21 (2000), 1253–1278.
- [33] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. 2000. On the Best Rank-1 and Rank-(R1,R2,...,RN) Approximation of Higher-Order Tensors. *SIAM J. Matrix Anal. Appl.* 21, 4 (March 2000), 1324–1342. <https://doi.org/10.1137/S0895479898346995>
- [34] Lieven De Lathauwer, Nico Vervliet, Martijn Bous, and Otto Debals. 2017. Dealing with curse and blessing of dimensionality through tensor decompositions.
- [35] Kaivalya M Dixit. 1991. The SPEC benchmarks. *Parallel computing* 17, 10-11 (1991), 1195–1209.
- [36] L. Eldn and B. Savas. 2009. A Newton-like Grassmann Method for Computing the Best Multilinear Rank-(r1,r2,...,rN) Approximation of a Tensor. *SIAM J. Matrix Anal. Appl.* 31, 2 (2009), 248–271. <https://doi.org/10.1137/070688316> [arXiv:https://doi.org/10.1137/070688316](http://arxiv.org/abs/https://doi.org/10.1137/070688316)
- [37] Evgeny Epifanovskiy, Michael Wormit, Tomasz Ku, Arie Landau, Dmitry Zuev, Kirill Khistyayev, Prashant Manohar, Ilya Kaliman, Andreas Dreuw, and Anna I. Krylov. [n. d.]. New implementation of high-level correlated methods using a general block tensor library for high-performance electronic structure calculations. *Journal of Computational Chemistry* 34, 26 ([n. d.]), 2293–2309. <https://doi.org/10.1002/jcc.23377> [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.23377](http://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.23377)
- [38] Glen Evenbly and Guifré Vidal. 2009. Algorithms for entanglement renormalization. *Physical Review B* 79, 14 (2009), 144108.
- [39] Xiaomin Fang and Rong Pan. 2014. Fast DTT: a near linear algorithm for decomposing a tensor into factor tensors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 967–976.
- [40] Vittorio Giovannetti, Simone Montangero, and Rosario Fazio. 2008. Quantum multiscale entanglement renormalization ansatz channels. *Physical review letters* 101, 18 (2008), 180503.
- [41] Alex A Gorodetsky, Sertac Karaman, and Youssef M Marzouk. [n. d.]. Efficient High-Dimensional Stochastic Optimal Motion Control using Tensor-Train Decomposition.
- [42] Lars Grasedyck. 2010. Hierarchical Singular Value Decomposition of Tensors. *SIAM J. Matrix Anal. Appl.* 31, 4 (May 2010), 2029–2054. <https://doi.org/10.1137/090764189>

- [43] Lars Grasedyck, Daniel Kressner, and Christine Tobler. 2013. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen* 36, 1 (2013), 53–78.
- [44] W. Hackbusch and S. Kühn. 2009. A New Scheme for the Tensor Representation. *Journal of Fourier Analysis and Applications* 15, 5 (01 Oct 2009), 706–722. <https://doi.org/10.1007/s00041-009-9094-9>
- [45] Samantha Hansen, Todd Plantenga, and Tamara G. Kolda. 2015. Newton-based optimization for Kullback-Leibler nonnegative tensor factorizations. *Optimization Methods and Software* 30 (2015), 1002–1029.
- [46] Richard Harshman and Margaret Lundy. 1996. Uniqueness proof for a family of models sharing features of Tucker’s three-mode factor analysis and PARAFAC/Candecomp. *Psychometrika* 61, 1 (1996), 133–154. <http://EconPapers.repec.org/RePEc:spr:psycho:v:61:y:1996:i:1:p:133-154>
- [47] R. A. Harshman. 1970. Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis. *UCLA Working Papers in Phonetics* 16, 1 (1970), 84.
- [48] R. A. Harshman. 1972. PARAFAC2: Mathematical and technical notes. *UCLA Working Papers in Phonetics* 22 (1972), 30–44.
- [49] Eric Hein, Tom Conte, Jeffrey S. Young, Srinivas Eswar, Jiajia Li, Patrick Lavin, Richard Vuduc, and Jason Riedy. 2018. An Initial Characterization of the Emu Chick. *2018 IEEE International Parallel and Distributed Processing Symposium Workshops* (May 2018), 10.
- [50] J. Henderson, J. C. Ho, A. N. Kho, J. C. Denny, B. A. Malin, J. Sun, and J. Ghosh. 2017. Granite: Diversified, Sparse Tensor Factorization for Electronic Health Record-Based Phenotyping. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 214–223. <https://doi.org/10.1109/ICHI.2017.61>
- [51] F. L. Hitchcock. 1927. The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys* 6, 1 (1927), 164–189.
- [52] Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. 2014. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of biomedical informatics* 52 (2014), 199–211.
- [53] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. 2014. Extracting Phenotypes from Patient Claim Records Using Nonnegative Tensor Factorization. In *Brain Informatics and Health*. Springer, 142–151.
- [54] Joyce C. Ho, Joydeep Ghosh, and Jimeng Sun. 2014. Marble: High-throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’14)*. ACM, New York, NY, USA, 115–124. <https://doi.org/10.1145/2623330.2623658>
- [55] F. Huang, N. Niranjan U., I. Perros, R. Chen, J. Sun, and A. Anandkumar. 2014. Scalable Latent Tree Model and its Application to Health Analytics. *ArXiv e-prints* (June 2014). arXiv:cs.LG/1406.4566
- [56] Brian Hutchinson, Li Deng, and Dong Yu. 2013. Tensor deep stacking networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35, 8 (2013), 1944–1957.
- [57] M. Ishteva, P. Absil, S. Van Huffel, and L. De Lathauwer. 2011. Best Low Multilinear Rank Approximation of Higher-Order Tensors, Based on the Riemannian Trust-Region Scheme. *SIAM J. Matrix Anal. Appl.* 32, 1 (2011), 115–135. <https://doi.org/10.1137/090764827> arXiv:https://doi.org/10.1137/090764827
- [58] Mariya Ishteva, Haesun Park, and Le Song. 2013. Unfolding Latent Tree Structures using 4th Order Tensors. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 316–324.
- [59] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. 2015. Generalization Bounds for Neural Networks through Tensor Factorization. *CoRR abs/1506.08473* (2015). <http://arxiv.org/abs/1506.08473>
- [60] Inah Jeon, Evangelos E. Papalexakis, and Christos Faloutsos U Kang. 2015. HaTen2: Billion-scale Tensor Decompositions (Version 1.0). Available from <http://datalab.snu.ac.kr/haten2/>.
- [61] Meng Jiang, Peng Cui, Fei Wang, Xinran Xu, Wenwu Zhu, and Shiqiang Yang. 2014. FEMA: flexible evolutionary multi-faceted analysis for dynamic behavioral pattern discovery. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1186–1195.
- [62] Tao Jiang and Nicholas D Sidiropoulos. 2004. Kruskal’s permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models with constant modulus constraints. *Signal Processing, IEEE Transactions on* 52, 9 (2004), 2625–2636.
- [63] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma,

- Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. 2017. In-Datcenter Performance Analysis of a Tensor Processing Unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA '17)*. ACM, New York, NY, USA, 1–12. <https://doi.org/10.1145/3079856.3080246>
- [64] Ilya A. Kaliman and Anna I. Krylov. [n. d.]. New algorithm for tensor contractions on multi-core CPUs, GPUs, and accelerators enables CCSD and EOM-CCSD calculations with over 1000 basis functions on a single compute node. *Journal of Computational Chemistry* 38, 11 ([n. d.]), 842–853. <https://doi.org/10.1002/jcc.24713> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.24713>
- [65] U. Kang, Evangelos Papalexakis, Abhay Harpale, and Christos Faloutsos. 2012. GigaTensor: Scaling Tensor Analysis Up by 100 Times - Algorithms and Discoveries. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, New York, NY, USA, 316–324. <https://doi.org/10.1145/2339530.2339583>
- [66] Arie Kapteyn, Heinz Neudecker, and Tom Wansbeek. 1986. An approach to n-mode components analysis. *Psychometrika* 51, 2 (01 Jun 1986), 269–275. <https://doi.org/10.1007/BF02293984>
- [67] Alexandros Karatzoglou, Xavier Amatriain, Linas Baltrunas, and Nuria Oliver. 2010. Multiverse Recommendation: N-dimensional Tensor Factorization for Context-aware Collaborative Filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*. ACM, New York, NY, USA, 79–86. <https://doi.org/10.1145/1864708.1864727>
- [68] Lars Karlsson, Daniel Kressner, and Andr Uschmajew. 2016. Parallel algorithms for tensor completion in the CP format. *Parallel Comput.* 57 (2016), 222 – 234. <https://doi.org/10.1016/j.parco.2015.10.002>
- [69] Oguz Kaya and Bora Uçar. 2015. Scalable Sparse Tensor Decompositions in Distributed Memory Systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '15)*. ACM, New York, NY, USA, Article 77, 11 pages. <https://doi.org/10.1145/2807591.2807624>
- [70] O. Kaya and B. Uar. 2018. Parallel Candecomp/Parafac Decomposition of Sparse Tensors Using Dimension Trees. *SIAM Journal on Scientific Computing* 40, 1 (2018), C99–C130. <https://doi.org/10.1137/16M1102744> arXiv:<https://doi.org/10.1137/16M1102744>
- [71] Venera Khoromskaia and Boris N Khoromskij. 2018. *Tensor Numerical Methods in Quantum Chemistry*. Walter de Gruyter GmbH & Co KG.
- [72] Henk A. L. Kiers and Albert der Kinderen. [n. d.]. A fast method for choosing the numbers of components in Tucker3 analysis. *Brit. J. Math. Statist. Psych.* 56, 1 ([n. d.]), 119–125. <https://doi.org/10.1348/000711003321645386> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1348/000711003321645386>
- [73] AJ KleinOowski and David J Lilja. 2002. MinneSPEC: A new SPEC benchmark workload for simulation-based computer architecture research. *IEEE Computer Architecture Letters* 1, 1 (2002), 7–7.
- [74] T. Kolda and B. Bader. 2009. Tensor Decompositions and Applications. *SIAM Rev.* 51, 3 (2009), 455–500. <https://doi.org/10.1137/07070111X> arXiv:<http://dx.doi.org/10.1137/07070111X>
- [75] Tamara G. Kolda and Brett W. Bader. 2006. The TOPHITS model for higher-order web link analysis. In *Workshop on link analysis, counterterrorism and security*, Vol. 7. 26–29.
- [76] Tamara G. Kolda and Jimeng Sun. 2008. Scalable Tensor Decompositions for Multi-aspect Data Mining. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM '08)*. IEEE Computer Society, Washington, DC, USA, 363–372. <https://doi.org/10.1109/ICDM.2008.89>
- [77] Christoph Kppl and Hans-Joachim Werner. 2016. Parallel and Low-Order Scaling Implementation of Hartree-Fock Exchange Using Local Density Fitting. *Journal of Chemical Theory and Computation* 12, 7 (2016), 3122–3134. <https://doi.org/10.1021/acs.jctc.6b00251> arXiv:<https://doi.org/10.1021/acs.jctc.6b00251> PMID: 27267488.
- [78] Charles-Francois V Latchoumane, Francois-Benois Vialatte, Jordi Solé-Casals, Monique Maurice, Sunil R Wimalaratna, Nigel Hudson, Jaeseung Jeong, and Andrzej Cichocki. 2012. Multiway array decomposition analysis of EEGs in Alzheimer's disease. *Journal of neuroscience methods* 207, 1 (2012), 41–50.
- [79] Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Oseledets, and Victor Lempitsky. 2014. Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition. *arXiv preprint arXiv:1412.6553* (2014).
- [80] Chunho Lee, Miodrag Potkonjak, and William H Mangione-Smith. 1997. MediaBench: a tool for evaluating and synthesizing multimedia and communications systems. In *Proceedings of the 30th annual ACM/IEEE international symposium on Microarchitecture*. IEEE Computer Society, 330–335.
- [81] Cannada A. Lewis, Justus A. Calvin, and Edward F. Valeev. 2016. Clustered Low-Rank Tensor Format: Introduction and Application to Fast Construction of Hartree-Fock Exchange. *Journal of Chemical Theory and Computation* 12, 12 (2016), 5868–5880. <https://doi.org/10.1021/acs.jctc.6b00884> arXiv:<https://doi.org/10.1021/acs.jctc.6b00884> PMID: 27783506.
- [82] Ang Li, Shuaiwen Leon Song, Jieyang Chen, Xu Liu, Nathan Tallent, and Kevin Barker. 2018. Tartan: Evaluating Modern GPU Interconnect via a Multi-GPU Benchmark Suite. In *2018 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 191–202.

- [83] Jiajia Li. 2018. *Scalable tensor decompositions in high performance computing environments*. Ph.D. Dissertation. Georgia Institute of Technology, Atlanta, GA, USA.
- [84] J. Li, J. Choi, I. Perros, J. Sun, and R. Vuduc. 2017. Model-Driven Sparse CP Decomposition for Higher-Order Tensors. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 1048–1057. <https://doi.org/10.1109/IPDPS.2017.80>
- [85] Jiajia Li, Yuchen Ma, Chenggang Yan, and Richard Vuduc. 2016. Optimizing Sparse Tensor Times Matrix on Multi-core and Many-core Architectures. In *Proceedings of the Sixth Workshop on Irregular Applications: Architectures and Algorithms (IA'3 '16)*. IEEE Press, Piscataway, NJ, USA, 26–33. <https://doi.org/10.1109/IA3.2016.10>
- [86] Jiajia Li, Jimeng Sun, and Richard Vuduc. 2018. HiCOO: Hierarchical storage of sparse tensors. In *Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC)*. Dallas, TX, USA. (to appear).
- [87] Jiajia Li, Guangming Tan, Mingyu Chen, and Ninghui Sun. 2013. SMAT: An Input Adaptive Auto-tuner for Sparse Matrix-vector Multiplication. In *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '13)*. ACM, New York, NY, USA, 117–126. <https://doi.org/10.1145/2491956.2462181>
- [88] Z. Li, A. Uschmajew, and S. Zhang. 2015. On Convergence of the Maximum Block Improvement Method. *SIAM Journal on Optimization* 25, 1 (2015), 210–233. <https://doi.org/10.1137/130939110> arXiv:<https://doi.org/10.1137/130939110>
- [89] B. Liu, C. Wen, A. D. Sarwate, and M. M. Dehnavi. 2017. A Unified Optimization Approach for Sparse Tensor Operations on GPUs. In *2017 IEEE International Conference on Cluster Computing (CLUSTER)*. 47–57. <https://doi.org/10.1109/CLUSTER.2017.75>
- [90] Yuchen Ma, Jiajia Li, Xiaolong Wu, Chenggang Yan, Jimeng Sun, and Richard Vuduc. 2018. Optimizing sparse tensor times matrix on GPUs. *J. Parallel and Distrib. Comput.* (2018). <https://doi.org/10.1016/j.jpdc.2018.07.018>
- [91] Samuel Manzer, Evgeny Epifanovsky, Anna I. Krylov, and Martin Head-Gordon. 2017. A General Sparse Tensor Framework for Electronic Structure Theory. *Journal of Chemical Theory and Computation* 13, 3 (2017), 1108–1116. <https://doi.org/10.1021/acs.jctc.6b00853> arXiv:<https://doi.org/10.1021/acs.jctc.6b00853> PMID: 28118011.
- [92] Yasuko Matsubara, Yasushi Sakurai, Willem G van Panhuis, and Christos Faloutsos. 2014. FUNNEL: automatic mining of spatially coevolving epidemics. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 105–114.
- [93] Martin J. Mohlenkamp. 2010. Musings on Multilinear Fitting.
- [94] M. Mørup, L. K. Hansen, S. M. Arnfred, L. Lim, and K. H. Madsen. 2008. Shift Invariant Multilinear Decomposition of Neuroimaging Data. *accepted for publication NeuroImage* 42, 4 (2008), 1439–50. <http://www2.imm.dtu.dk/pubdb/p.php?5551>
- [95] Naoki Nakatani and Garnet Kin-Lic Chan. 2013. Efficient tree tensor network states (TTNS) for quantum chemistry: Generalizations of the density matrix renormalization group algorithm. *The Journal of Chemical Physics* 138, 13 (2013), 134113. <https://doi.org/10.1063/1.4798639> arXiv:<https://doi.org/10.1063/1.4798639>
- [96] Alexander Novikov, Pavel Izmailov, Valentin Khrulkov, Michael Figurnov, and Ivan V. Oseledets. 2018. Tensor Train decomposition on TensorFlow (T3F). *CoRR abs/1801.01928* (2018). arXiv:1801.01928 <http://arxiv.org/abs/1801.01928>
- [97] Alexander Novikov, Dmitry Podoprikin, Anton Osokin, and Dmitry Vetrov. 2015. Tensorizing Neural Networks. *CoRR abs/1509.06569* (2015).
- [98] Alexander Novikov, Anton Rodomanov, Anton Osokin, and Dmitry Vetrov. 2014. Putting MRFs on a Tensor Train. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, Tony Jebara and Eric P. Xing (Eds.). JMLR Workshop and Conference Proceedings, 811–819. <http://jmlr.org/proceedings/papers/v32/novikov14.pdf>
- [99] HyungSeon Oh. 2016. Tensors in Power System Computation I: Distributed Computation for Optimal Power Flow, DC OPF. *CoRR abs/1605.06735* (2016).
- [100] Román Orús. 2014. A practical introduction to tensor networks: Matrix product states and projected entangled pair states. *Annals of Physics* 349 (2014), 117–158.
- [101] I. V. Oseledets. 2011. Tensor-Train Decomposition. *SIAM J. Sci. Comput.* 33, 5 (Sept. 2011), 2295–2317. <https://doi.org/10.1137/090752286>
- [102] Evangelos E. Papalexakis, Leman Akoglu, and Dino Ienco. 2013. Do more views of a graph help? Community detection and clustering in multi-graphs. In *Proceedings of the 16th International Conference on Information Fusion, FUSION 2013, Istanbul, Turkey, July 9-12, 2013*. 899–905. <http://ieeexplore.ieee.org/xpl/freeabs.all.jsp?arnumber=6641090>
- [103] Evangelos E. Papalexakis, Christos Faloutsos, Tom M. Mitchell, Partha Pratim Talukdar, Nicholas D. Sidiropoulos, and Brian Murphy. [n. d.]. *Turbo-SMT: Accelerating Coupled Sparse Matrix-Tensor Factorizations by 200x*. Chapter 14, 118–126. <https://doi.org/10.1137/1.9781611973440.14> arXiv:<http://epubs.siam.org/doi/pdf/10.1137/1.9781611973440.14>
- [104] Evangelos E. Papalexakis, Christos Faloutsos, and Nicholas D. Sidiropoulos. 2012. ParCube: Sparse Parallelizable Tensor Decompositions. In *Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I (ECML PKDD'12)*. Springer-Verlag, Berlin, Heidelberg, 521–536. https://doi.org/10.1007/978-3-642-33460-3_39

- [105] Evangelos E. Papalexakis, Christos Faloutsos, and Nicholas D. Sidiropoulos. 2015. ParCube: Sparse Parallelizable CANDECOMP-PARAFAC Tensor Decomposition. *ACM Trans. Knowl. Discov. Data* 10, 1, Article 3 (July 2015), 25 pages. <https://doi.org/10.1145/2729980>
- [106] Evangelos E Papalexakis and Nicholas D Sidiropoulos. 2011. Co-clustering as multilinear decomposition with sparse latent factors. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2064–2067.
- [107] Chong Peng, Justus A. Calvin, Fabijan Pavoevifj, Jinmei Zhang, and Edward F. Valeev. 2016. Massively Parallel Implementation of Explicitly Correlated Coupled-Cluster Singles and Doubles Using TiledArray Framework. *The Journal of Physical Chemistry A* 120, 51 (2016), 10231–10244. <https://doi.org/10.1021/acs.jpca.6b10150> arXiv:<https://doi.org/10.1021/acs.jpca.6b10150> PMID: 27966947.
- [108] Ioakeim Perros, Robert Chen, Richard Vuduc, and Jimeng Sun. 2015. Sparse Hierarchical Tucker Factorization and Its Application to Healthcare. In *Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM) (ICDM '15)*. IEEE Computer Society, Washington, DC, USA, 943–948. <https://doi.org/10.1109/ICDM.2015.29>
- [109] Ioakeim Perros, Evangelos E. Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. 2017. SPARTan: Scalable PARAFAC2 for Large & Sparse Data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. ACM, New York, NY, USA, 375–384. <https://doi.org/10.1145/3097983.3098014>
- [110] Eric T. Phipps and Tamara G. Kolda. 2018. Software for Sparse Tensor Decomposition on Emerging Computing Architectures. *CoRR* abs/1809.09175 (2018). arXiv:1809.09175 <http://arxiv.org/abs/1809.09175>
- [111] Jason A Poovey, Thomas M Conte, Markus Levy, and Shay Gal-On. 2009. A benchmark characterization of the EEMBC benchmark suite. *IEEE micro* 29, 5 (2009).
- [112] M. Rajih and P. Comon. 2005. Enhanced Line Search: A novel method to accelerate Parafac. In *2005 13th European Signal Processing Conference*. 1–4.
- [113] Niranjay Ravindran, Nicholas D. Sidiropoulos, Shaden Smith, and George Karypis. 2014. Memory-Efficient Parallel Computation of Tensor and Matrix Products for Big Tensor Decompositions. *Proceedings of the Asilomar Conference on Signals, Systems, and Computers* (2014).
- [114] Steffen Rendle, Leandro Balby Marinho, Alexandros Nanopoulos, and Lars Schmidt-Thieme. 2009. Learning Optimal Ranking with Tensor Factorization for Tag Recommendation. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. ACM, New York, NY, USA, 727–736. <https://doi.org/10.1145/1557019.1557100>
- [115] M. Reynolds, A. Doostan, and G. Beylkin. 2016. Randomized Alternating Least Squares for Canonical Tensor Decompositions: Application to A PDE With Random Data. *SIAM Journal on Scientific Computing* 38, 5 (2016), A2634–A2664. <https://doi.org/10.1137/15M1042802> arXiv:<https://doi.org/10.1137/15M1042802>
- [116] Bernardino Romera-Paredes, Min Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. 2013. Multilinear Multitask Learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML '13)*. JMLR.org, III–1444–III–1452. <http://dl.acm.org/citation.cfm?id=3042817.3043098>
- [117] B. Savas and L. Lim. 2010. Quasi-Newton Methods on Grassmannians and Multilinear Approximations of Tensors. *SIAM Journal on Scientific Computing* 32, 6 (2010), 3352–3393. <https://doi.org/10.1137/090763172> arXiv:<https://doi.org/10.1137/090763172>
- [118] Aaron Schein, John Paisley, David M. Blei, and Hanna Wallach. 2015. Bayesian Poisson Tensor Factorization for Inferring Multilateral Relations from Sparse Dyadic Event Counts. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1045–1054. <https://doi.org/10.1145/2783258.2783414>
- [119] Naser Sedaghati, Te Mu, Louis-Noel Pouchet, Srinivasan Parthasarathy, and P. Sadayappan. 2015. Automatic Selection of Sparse Matrix Representation on GPUs. In *Proceedings of the 29th ACM on International Conference on Supercomputing (ICS '15)*. ACM, New York, NY, USA, 99–108. <https://doi.org/10.1145/2751205.2751244>
- [120] Hendra Setiawan, Zhongqiang Huang, Jacob Devlin, Thomas Lamar, Rabih Zbib, Richard M. Schwartz, and John Makhoul. 2015. Statistical Machine Translation Features with Multitask Tensor Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. 31–41. <http://aclweb.org/anthology/P/P15/P15-1004.pdf>
- [121] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. 2017. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Transactions on Signal Processing* 65, 13 (July 2017), 3551–3582. <https://doi.org/10.1109/TSP.2017.2690524>
- [122] Nicholas D Sidiropoulos, Georgios B Giannakis, and Rasmus Bro. 2000. Blind PARAFAC receivers for DS-CDMA systems. *Signal Processing, IEEE Transactions on* 48, 3 (2000), 810–823.

- [123] Marco Signoretto, Quoc TranDinh, Lieven De Lathauwer, and Johan A. K. Suykens. 2014. Learning with tensors: a framework based on convex optimization and spectral regularization. *Machine Learning* 94, 3 (01 Mar 2014), 303–351. <https://doi.org/10.1007/s10994-013-5366-3>
- [124] Shaden Smith, Alec Beri, and George Karypis. 2017. Constrained Tensor Factorization with Accelerated AO-ADMM. In *46th International Conference on Parallel Processing (ICPP '17)*. IEEE.
- [125] Shaden Smith, Jee W. Choi, Jiajia Li, Richard Vuduc, Jongsoo Park, Xing Liu, and George Karypis. 2017. FROSTT: The Formidable Repository of Open Sparse Tensors and Tools. <http://frostt.io/>
- [126] Shaden Smith and George Karypis. 2016. A Medium-Grained Algorithm for Distributed Sparse Tensor Factorization. In *Parallel and Distributed Processing Symposium (IPDPS), 2016 IEEE International*. IEEE.
- [127] Shaden Smith and George Karypis. 2017. Accelerating the Tucker Decomposition with Compressed Sparse Tensors. In *European Conference on Parallel Processing*. Springer.
- [128] Shaden Smith, Jongsoo Park, and George Karypis. 2016. An Exploration of Optimization Algorithms for High Performance Tensor Completion. *Proceedings of the 2016 ACM/IEEE conference on Supercomputing* (2016).
- [129] Shaden Smith, Jongsoo Park, and George Karypis. 2017. Sparse Tensor Factorization on Many-Core Processors with High-Bandwidth Memory. *31st IEEE International Parallel & Distributed Processing Symposium (IPDPS'17)* (2017).
- [130] Shaden Smith, Niranjay Ravindran, Nicholas Sidiropoulos, and George Karypis. 2015. SPLATT: Efficient and Parallel Sparse Tensor-Matrix Multiplication. In *Proceedings of the 29th IEEE International Parallel & Distributed Processing Symposium (IPDPS)*.
- [131] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*. 926–934.
- [132] E. Solomonik and T. Hoefer. 2015. Sparse Tensor Algebra as a Parallel Programming Model. *ArXiv e-prints* (Nov. 2015). arXiv:cs.MS/1512.00066
- [133] Chenchen Song and Todd J. Martinez. 2016. Atomic orbital-based SOS-MP2 with tensor hypercontraction. I. GPU-based tensor construction and exploiting sparsity. *The Journal of Chemical Physics* 144, 17 (2016), 174111. <https://doi.org/10.1063/1.4948438> arXiv:https://doi.org/10.1063/1.4948438
- [134] Hyun Ah Song, Bryan Hooi, Marko Jereminov, Amritanshu Pandey, Lawrence T. Pileggi, and Christos Faloutsos. 2017. PowerCast: Mining and Forecasting Power Grid Sequences. In *ECML/PKDD*.
- [135] Le Song, Mariya Ishteva, Ankur Parikh, Eric Xing, and Haesun Park. 2013. Hierarchical Tensor Decomposition of Latent Tree Graphical Models. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 334–342.
- [136] Zhao Song, David P. Woodruff, and Huan Zhang. 2016. Sublinear Time Orthogonal Tensor Decomposition. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 793–801. <http://dl.acm.org/citation.cfm?id=3157096.3157185>
- [137] Laurent Sorber, Ignat Domanov, Marc Barel, and Lieven Lathauwer. 2016. Exact Line and Plane Search for Tensor Optimization. *Comput. Optim. Appl.* 63, 1 (Jan. 2016), 121–142. <https://doi.org/10.1007/s10589-015-9761-5>
- [138] L. Sorber, M. Van Barel, and L. De Lathauwer. 2013. Optimization-Based Algorithms for Tensor Decompositions: Canonical Polyadic Decomposition, Decomposition in Rank- $(L_r, L_r, 1)$ Terms, and a New Generalization. *SIAM Journal on Optimization* 23, 2 (2013), 695–720. <https://doi.org/10.1137/120868323> arXiv:https://doi.org/10.1137/120868323
- [139] Bor-Yiing Su and Kurt Keutzer. 2012. clSpMV: A Cross-Platform OpenCL SpMV Framework on GPUs. In *Proceedings of the 26th ACM International Conference on Supercomputing (ICS '12)*. ACM, New York, NY, USA, 353–364. <https://doi.org/10.1145/2304576.2304624>
- [140] Jimeng Sun, Spiros Papadimitriou, Ching-Yung Lin, Nan Cao, Shixia Liu, and Weihong Qian. 2009. MultiVis: Content-Based Social Network Exploration through Multi-way Visual Analysis. In *SDM*, Vol. 9. SIAM, 1063–1074.
- [141] Jimeng Sun, Dacheng Tao, and Christos Faloutsos. 2006. Beyond Streams and Graphs: Dynamic Tensor Analysis. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. ACM, New York, NY, USA, 374–383. <https://doi.org/10.1145/1150402.1150445>
- [142] Jian-Tao Sun, Hua-Jun Zeng, Huan Liu, Yuchang Lu, and Zheng Chen. 2005. CubeSVD: A Novel Approach to Personalized Web Search. In *Proceedings of the 14th International Conference on World Wide Web (WWW '05)*. ACM, New York, NY, USA, 382–390. <https://doi.org/10.1145/1060745.1060803>
- [143] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2008. Tag Recommendations Based on Tensor Dimensionality Reduction. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*. ACM, New York, NY, USA, 43–50. <https://doi.org/10.1145/1454008.1454017>
- [144] Dacheng Tao, Xuelong Li, Xindong Wu, Weiming Hu, and Stephen J. Maybank. 2007. Supervised Tensor Learning. *Knowl. Inf. Syst.* 13, 1 (Sept. 2007), 1–42. <https://doi.org/10.1007/s10115-006-0050-6>
- [145] Giorgio Tomasi and Rasmus Bro. 2006. A Comparison of Algorithms for Fitting the PARAFAC Model. *Comput. Stat. Data Anal.* 50, 7 (April 2006), 1700–1734. <https://doi.org/10.1016/j.csda.2004.11.013>

- [146] Ledyard R. Tucker. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31, 3 (01 Sep 1966), 279–311. <https://doi.org/10.1007/BF02289464>
- [147] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer. 2016. Tensorlab (Version 3.0). Available from <http://www.tensorlab.net>.
- [148] N. Vervliet and L. De Lathauwer. 2016. A Randomized Block Sampling Approach to Canonical Polyadic Decomposition of Large-Scale Tensors. *IEEE Journal of Selected Topics in Signal Processing* 10, 2 (March 2016), 284–295. <https://doi.org/10.1109/JSTSP.2015.2503260>
- [149] Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. 2014. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 145–154.
- [150] Haobin Wang and Michael Thoss. 2003. Multilayer formulation of the multiconfiguration time-dependent Hartree theory. *The Journal of Chemical Physics* 119, 3 (2003), 1289–1299. <https://doi.org/10.1063/1.1580111> arXiv:<https://doi.org/10.1063/1.1580111>
- [151] L. Wang, J. Zhan, C. Luo, Y. Zhu, Q. Yang, Y. He, W. Gao, Z. Jia, Y. Shi, S. Zhang, C. Zheng, G. Lu, K. Zhan, X. Li, and B. Qiu. 2014. BigDataBench: A big data benchmark suite from internet services. In *2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*, 488–499. <https://doi.org/10.1109/HPCA.2014.6835958>
- [152] Yichen Wang, Robert Chen, Joydeep Ghosh, Joshua C. Denny, Abel Kho, You Chen, Bradley A. Malin, and Jimeng Sun. 2015. Rubik: Knowledge Guided Tensor Factorization and Completion for Health Data Analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1265–1274. <https://doi.org/10.1145/2783258.2783395>
- [153] Kishan Wimalawarne, Masashi Sugiyama, and Ryota Tomioka. 2014. Multitask learning meets tensor factorization: task imputation via convex optimization. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2825–2833. <http://papers.nips.cc/paper/5628-multitask-learning-meets-tensor-factorization-task-imputation-via-convex-optimization.pdf>
- [154] Stephen Wright and Jorge Nocedal. 1999. Numerical optimization. *Springer Science* 35, 67–68 (1999), 7.
- [155] Yanfei Xu, Liang Zhang, and Wei Liu. 2006. Cubic Analysis of Social Bookmarking for Personalized Recommendation. In *Frontiers of WWW Research and Development - APWeb 2006*, Xiaofang Zhou, Jianzhong Li, HengTao Shen, Masaru Kitsuregawa, and Yanchun Zhang (Eds.). Lecture Notes in Computer Science, Vol. 3841. Springer Berlin Heidelberg, 733–738. <https://doi.org/10.1007/11610113.66>
- [156] Dong Yu, Li Deng, and Frank Seide. 2012. Large Vocabulary Speech Recognition Using Deep Tensor Neural Networks.. In *INTERSPEECH*.
- [157] Qi Rose Yu and Yan Liu. 2016. Learning from Multiway Data: Simple and Efficient Tensor Regression. *CoRR* abs/1607.02535 (2016). arXiv:1607.02535 <http://arxiv.org/abs/1607.02535>
- [158] Rose Yu, Guangyu Li, and Yan Liu. 2017. Tensor Regression Meets Gaussian Processes. *CoRR* abs/1710.11345 (2017). arXiv:1710.11345 <http://arxiv.org/abs/1710.11345>
- [159] Rose Yu, Stephan Zheng, Anima Anandkumar, and Yisong Yue. 2018. Long-term Forecasting using Tensor-Train RNNs. <https://openreview.net/forum?id=HJJ0w--0W>
- [160] Zheng Zhang, Kim Batselier, Haotian Liu, Luca Daniel, and Ngai Wong. 2017. Tensor computation: A new framework for high-dimensional problems in EDA. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 36, 4 (2017), 521–536.
- [161] Zheng Zhang, Xiu Yang, Ivan V Oseledets, George E Karniadakis, and Luca Daniel. 2015. Enabling high-dimensional hierarchical uncertainty quantification by ANOVA and tensor-train decomposition. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34, 1 (2015), 63–76.
- [162] Qibin Zhao, Cesar F Caiafa, Danilo P. Mandic, Liqing Zhang, Tonio Ball, Andreas Schulze-bonhage, and Andrzej S. Cichocki. 2011. Multilinear Subspace Regression: An Orthogonal Tensor Decomposition Approach. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1269–1277. <http://papers.nips.cc/paper/4328-multilinear-subspace-regression-an-orthogonal-tensor-decomposition-approach.pdf>
- [163] Yue Zhao, Jiajia Li, Chunhua Liao, and Xipeng Shen. 2018. Bridging the Gap Between Deep Learning and Sparse Matrix Format Selection. In *Proceedings of the 23rd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '18)*. ACM, New York, NY, USA, 94–108. <https://doi.org/10.1145/3178487.3178495>
- [164] Guoxu Zhou, Andrzej Cichocki, and Shengli Xie. 2014. Decomposition of Big Tensors With Low Multilinear Rank. *CoRR* abs/1412.1885 (2014).
- [165] Hua Zhou, Lexin Li, and Hongtu Zhu. 2013. Tensor Regression with Applications in Neuroimaging Data Analysis. *J. Amer. Statist. Assoc.* 108, 502 (2013), 540–552. <https://doi.org/10.1080/01621459.2013.776499> arXiv:<https://doi.org/10.1080/01621459.2013.776499> PMID: 24791032.