Performance of Dense and Sparse Tensor Operations in Applications

Data Compression with Low-Rank Tensor Methods.

Melven.Roehrig-Zoellner@DLR.de

Melven Röhrig-Zöllner*, Dr. Jonas Thies†, Dr. Achim Basermann*, Prof. Dr. Axel Klawonn‡

* Institute for Software Technology, German Aerospace Center (DLR), Cologne, Germany
† Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands
‡ Department of Mathematics and Computer Science, University of Cologne, Cologne, Germany

MOTIVATION

- Data centers emit about 1% of global CO_2 emissions.[1] (for comparison: aviation emits about 2%)
- Need to process larger and larger data...

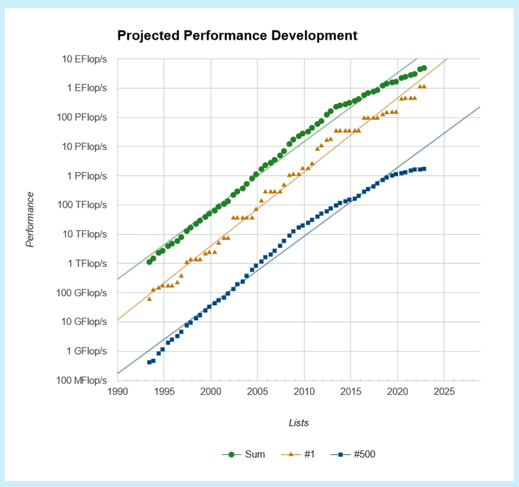
Goal: Investigate new algorithms based on compressed data.

METHODS

- Performance modelling: combine hardware abstractions with characteristics of an algorithm ⇒ Identify bottlenecks/optimizations.[2]
- Co-design of low-level building blocks and numerical algorithms: (Often) Faster through more work and less data transfers.
- Improve data science algorithms (clustering, medical image analysis).

HPC SYSTEMS

- Increasing overall performance BUT
  - more levels of parallelism
  - 'slow' data transfers

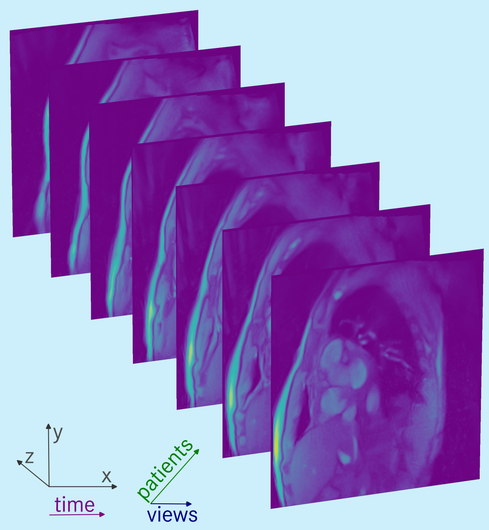


(Source: https://top500.org/)

+

MULTI-DIMENSIONAL DATA

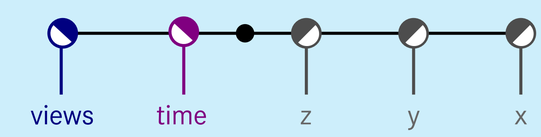
- Scientific data often well structured
- Example: real-time MRT images



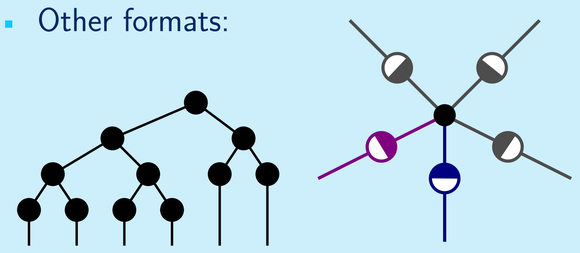
+

TENSOR-NETWORKS

- Inspired by physics: 'few' parameters for expontial large quantum state
- Extend singular value decompositions (SVD) to higher dimensions
- Tensor-Train[3] format:



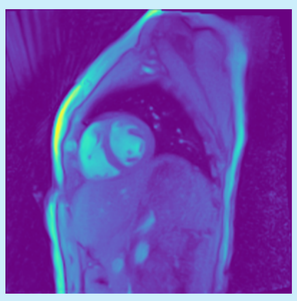
- Other formats:



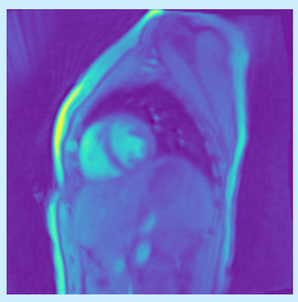
⇓

COMPRESSED DATA

- Whole data set: 124 MB vs. 16.3 GB
- Main features still visible



(original)



(compressed)

→

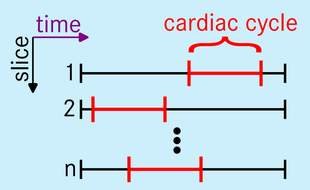
NUMERICAL ALGORITHMS

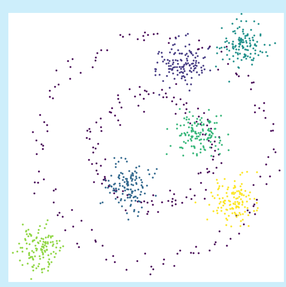
- Linear algebra with compressed data
  - additions, scalar products, linear operators
  - linear solvers, eigensolvers
- Black box approximation

→

FASTER DATA ANALYSIS

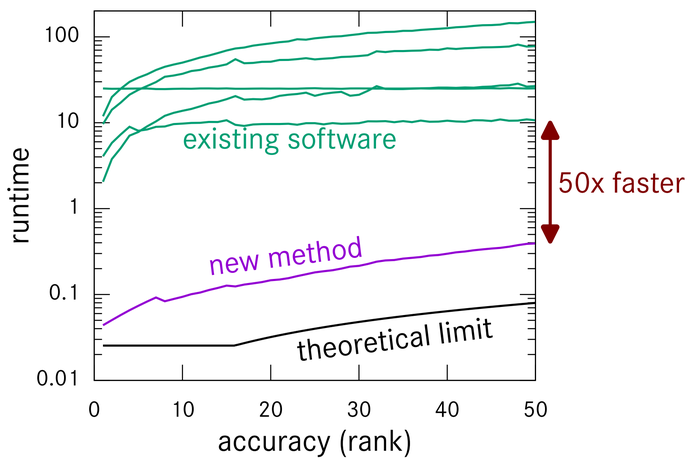
- Preprocessing of real-time MRT data (select matching frames in all slices)
- Approximate spectral clustering





RESULTS

- Faster Tensor-Train compression[4]
  - Ideal runtime estimated with performance model
  - Optimized algorithm & implementation
- Faster Tensor-Train arithmetic & solvers
  - Exploit structure to avoid operations
  - work-in-progress (estimated ∼2x speedup)



FUTURE WORK

- Complete 'compressed' workflow
- Interaction of neural networks and tensor networks (training/inference with compressed data)
- Further algorithms for data science with different tensor network formats

REFERENCES

1. Kamiya, G. *et al. Data Centres and Data Transmission Networks.* (International Energy Agency (IEA), 2022).
2. Williams, S. *et al.* Roofline: An Insightful Visual Performance Model for Multicore Architectures. *Communications of the ACM* **52,** 65–76 (2009).
3. Oseledets, I. V. Tensor-Train Decomposition. *SIAM Journal on Scientific Computing* **33,** 2295–2317 (2011).
4. Röhrig-Zöllner, M., Thies, J. & Basermann, A. Performance of the Low-Rank TT-SVD for Large Dense Tensors on Modern MultiCore CPUs. *SIAM Journal on Scientific Computing* **44,** C287–C309 (July 2022).