Title

Can tile low-rank compression live up to expectations? An application to 3D multidimensional deconvolution

Authors

Hong, Yuxi

Ravasi, Matteo

Ltaief, Hatem

et al.

Publication Date

2023-12-14

DOI

10.1190/image2023-3906829.1

Peer reviewed

Can tile low-rank compression live up to expectations? An application to 3D multi-dimensional deconvolution

Yuxi Hong, Matteo Ravasi, Hatem Ltaief, David Keyes, KAUST*

SUMMARY

Wave-equation-based seismic processing algorithms have been developed over the years with the aim of handling the 3D, full-wavefield nature of seismic waves. Multi-Dimensional Deconvolution (MDD) is one of such algorithms, commonly used to remove overburden-related effects from up/down separated wavefields (e.g., removal of free-surface multiples from ocean-bottom data). However, MDD comes with several computational challenges; this is especially the case for its time-domain implementation, which requires repeated access to Terabyte-scale seismic datasets. In this work, we present a novel algorithmic solution that leverages the inherent data sparsity of seismic data in the frequency domain by means of tile low-rank data compression. We further rely on so-called Hilbert reordering to achieve a boost in the compressibility of the dataset under study. Tile Low-Rank Matrix Vector Multiplication (TLR-MVM) is then introduced to speed up the Multi-Dimensional Convolution (MDC) operator that lies at the core of the MDD algorithm. The presented solution is tested on a realistic 3D seismic dataset modelled from the SEG/EAGE Overthrust model, and the impact of two key parameters in tile low-rank compression algorithm, namely tile size and error accuracy, is thoroughly investigated. Inversion is finally performed using the LSQR solver with all MDC operations performed onto GPUs. On a 4 A100 cluster, successful deconvolution for single virtual source is accomplished within 2 minutes (including I/O). To conclude, the proposed algorithm is deployed onto several mainstream hardware the associated roofline performance model is presented.

INTRODUCTION

Traditional algorithms in reflection seismology rely on strong assumptions about wave propagation in the subsurface, typically considering a 1D layered medium assumption. As a result, these algorithms overlook the propagation effects caused by lateral heterogeneities. In the 1990s, various wave-equation-based processing methods emerged in an attempt to handle the multi-dimensional nature of seismic waves (Verschuur, 1992; Jakubowicz, 1998; Amundsen, 2001). These methods have been later reformulated as inverse problems (van Groenestijn and Verschuur, 2009; Lopez and Verschuur, 2015; Wapenaar et al., 2011, 2014), offering superior processing capabilities for enhancing the imaging of seismic data acquired in complex geological settings.

Despite their undoubted potential, inversion-based methods present significant computational challenges, mainly due to the need for repeated access to the entire seismic dataset when solving the associated inverse problem (Ravasi and Vasconcelos, 2021). Consequently, the adoption of these techniques in industrial applications is still in its early stages, as it requires careful consideration and optimization to address the associated computational demands.

Multi-Dimensional Deconvolution (MDD) is a technique that can be used to remove free-surface effects from ocean-bottom seismic data (Wapenaar et al., 2011; Ravasi et al., 2015; Boiero and Bagaini, 2020; Ravasi et al., 2022b) or overburden-related multiples from synthesised data at a target depth of interest (van der Neut and Herrmann, 2013; Vasconcelos et al., 2017; Vargas et al., 2021). In its time-domain implementation, which we will consider in this work, the overall computational cost of MDD is dominated by the repeated application of the so-called Multi-Dimensional Convolution (MDC) operator and its adjoint. In this work, we assess the feasibility of applying tile low-rank compressed MDC on a realistic 3D synthetic dataset. Our findings show that, provided a proper sorting is applied to sources and receivers in the kernel of the MDC operator, frequency-domain seismic data can be significantly compressed. This ultimately leads to a faster and less memory demanding MDD process. This corroborates our previous findings in the context of Marchenko-based redatuming for a synthetic dataset modelled in a much simpler geological setting (Hong et al., 2021; Ravasi et al., 2022a).

Our contribution is three fold,

- We create a realistic 3D seismic dataset and assess for the first time the benefit of applying tile low-rank compression to its frequency matrices.
- We show the importance of applying distance-aware reordering to improve the compressibility of seismic data.
- We present a GPU-friendly implementation of MDD that scales up to 4 A100 GPUs. Future work will extend our algorithm to multiple virtual sources, i.e., implementing a tile low-rank version of the matrix-matrix multiplication.

METHOD

MDC in a nutshell

At the core of the MDD algorithm lies a linear operator performing batch matrix-vector multiplication with the frequency matrices of the kernel of the MDC operator (here, the down-going wavefield). This operator can be written in a compact matrix-vector form

$$\mathbf{y} = \mathbf{F}^H \mathbf{K} \mathbf{F} \mathbf{x}, \tag{1}$$

where $\mathbf{F}$ and $\mathbf{F}^H$ represent forward and inverse Fast Fourier Transform applied along the time/frequency axes (implemented as subroutines, not dense matrix multiplications), and $\mathbf{x}$ and $\mathbf{y}$ contain vectorized versions of the input and output functions. Finally, $\mathbf{K}$ is the operator that performs repeated Matrix-Vector

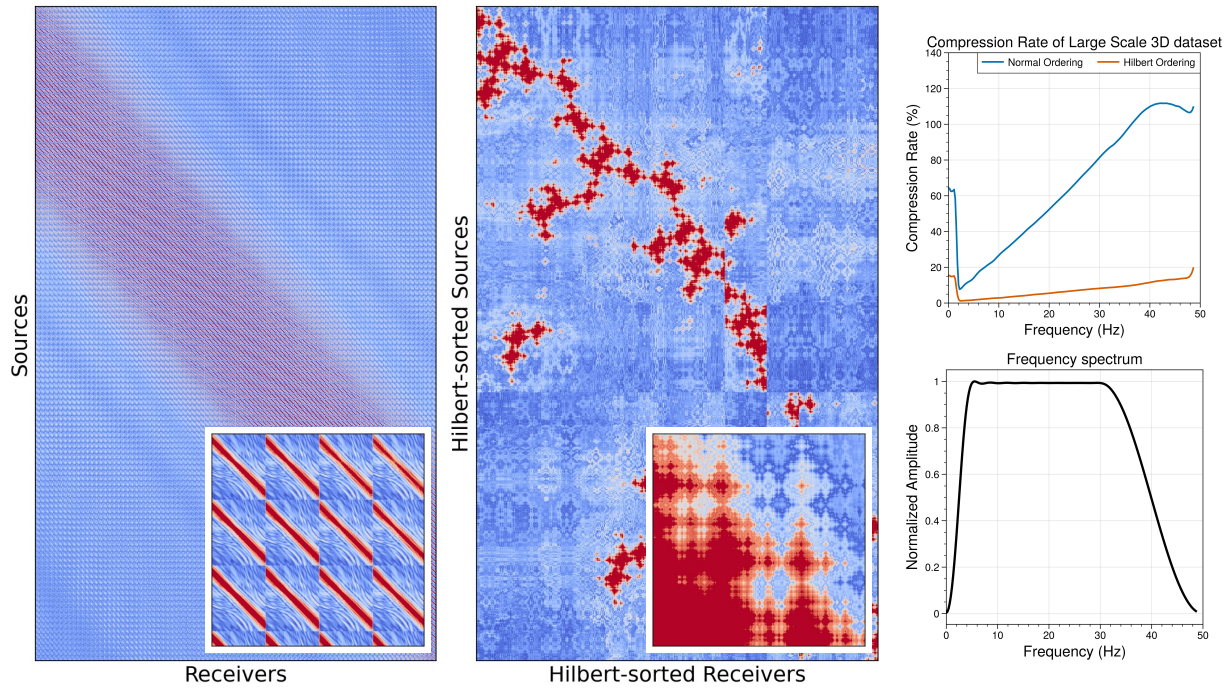

Figure 1: Visualization of a Single Frequency Matrix in Normal Ordering (left) and Hilbert Ordering (middle), The Compressed Rate and Frequency Spectrum (right).

Multiplications (MVMs) with the frequency matrices belonging to the seismic bandwidth of interest.

When dealing with large-scale 3D datasets, this operation dominates the cost of MDD mostly because of its heavy memory requirements. Recent research has shown that this challenge can be mitigated by performing a pre-proprocessing step where the frequency matrices are compressed in a tile low-rank fashion, i.e., the matrix is divided into tiles and a singular value decomposition (SVD) is applied to each tile separately to capture the most significant information. This allows identifying local low-rankness even though the matrix may not be globally low-rank. After applying SVD to tile, we increasingly add singular values to the approximated matrix until the norm of residual matrix is smaller than a percent of original one defined by the user. The resulting singular vectors, usually referred to as U and V bases, are then stored on disk for subsequent computations.

Hilbert reordering of frequency matrices

Before performing matrix-vector multiplication, one has the flexibility to re-arrange the rows (sources) and columns (receivers) of the matrix as long as the input and output vectors are re-arranged accordingly. This is also the case for the TLR-MVM version of this operation. Such a re-ordering can be applied prior to tiling and SVD with the aim of increasing the compressibility of the resulting tiles. The approach adopted here involves using the Hilbert reordering method, which allows us to rearrange sources and receivers based on their geographical distance rather than their natural or cable-based ordering. Subsequently, when applying the MDC operator, we employ a tiled batch matrix-vector multiplication approach, resulting in faster computations due to the smaller size of the singular vectors compared to the dense matrix.

TLR-MVM GPU implementation

We now briefly recall the GPU TLR-MVM implementation used within our MDC operator. Ltaief et al. (2021) first introduced TLR-MVM in the context of hard real-time controllers for ground-based telescopes. It operates on the pre-computed U/V bases of matrix, which are stacked together to increase memory alignment. TLR-MVM contains three phases: the first phase involves a batched GEMV with variable size U bases; The second phase is an element-wise reshuffling, and The third phase is also a batched GEMV with variable size V bases. The TLR-MVM implementation in Hong et al. (2021) is done in C++ on the NEC SX-Aurora TSUBASA vector engine using the MPI+OpenMP programming models. To enhance performance and reduce idle times during the simultaneous processing of all seismic frequency matrices in single complex precision arithmetic, a load balancing technique is introduced. The objective here is to extend TLR-MVM to a broader range of systems, including x86, ARM, and GPU platforms.

For x86 and ARM systems, MPI+OpenMP is natively supported, and achieving portability requires moderate effort. However, to maximize performance, careful mapping of MPI processes and corresponding OpenMP threads is necessary to align with the underlying core/socket packaging. In the case of GPU implementation (specifically NVIDIA GPUs in this study), MPI

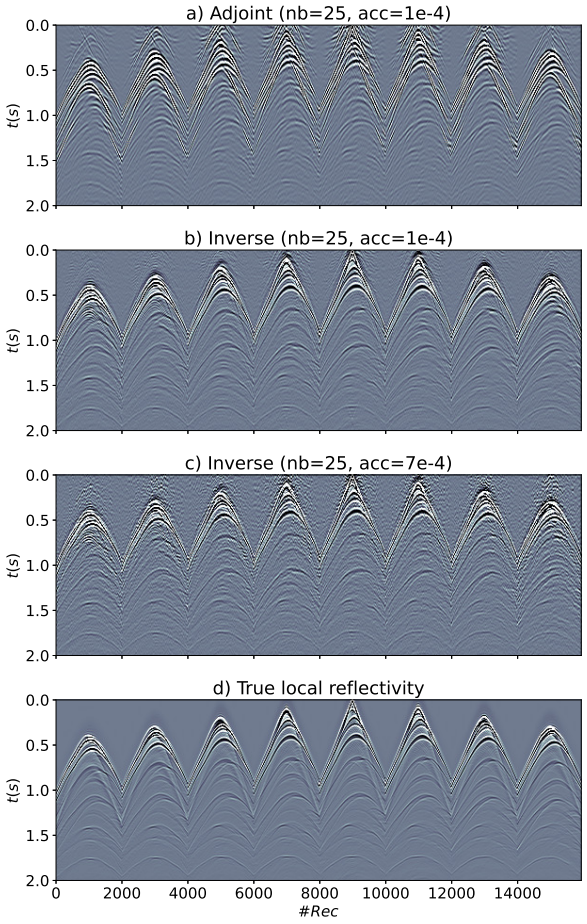

Figure 2: MDD results. a) Cross-correlation (i.e., adjoint), b) and c) inversion with 30 iterations of LSQR and MDC kernels compressed with different accuracy, and d) true local reflectivity. Each panel shows the data from 8 equally spaced receiver lines for a virtual source in the middle of the receiver grid.

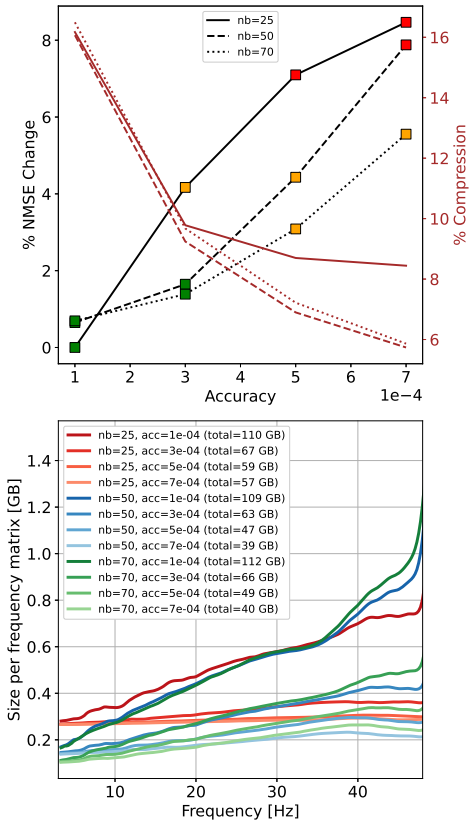

Figure 3: Effect of compression on the quality of the MDD inversion product. Top: Black lines represent percentage change of normalized mean square error of each solution against the benchmark solution with $nb = 25$ and $acc = 1e-4$. Brown lines refer to percentage of compression of each approximation compared to the original, dense solution. Bottom: Aggregated size of the U and V bases as function of frequency for the different combinations of nb and acc.

serves as the communication bridge across GPUs, while the CUDA software ecosystem is utilized within the GPU. Streams are employed to launch batch MVMs with varying sizes and monitor data dependencies during the TLR-MVM computation. Leveraging the CUDA Graph framework allows for asynchronous execution of these streams, reducing kernel launch overheads associated with small data structures. A total of 20 streams are used to optimize the performance.

NUMERICAL EXAMPLES

To begin with, we describe the 3D seismic dataset used in this study. The SEG/EAGE Overthrust model, Aminzadeh et al. (1997) is modified by including a 300m water column in order to mimic an ocean-bottom acquisition scenario. The modelled dataset is composed of a grid of 217x120 sources and 177x90 receivers with 20m spacing in both inline and crosslink directions. Pressure and particle velocity data are modeled with a flat wavelet up to 45Hz for a total time of 4.5sec, with each dataset having an effective size of 1.7TB. After wavefield separation, the downgoing pressure data is transformed in the frequency domain.

Each frequency matrix is then compressed using the tile low-rank compression algorithm with tiles of size 256x256. Figure 1 shows that the arrangement of rows and columns plays a crucial role in the compressibility of such matrices. As Hilbert reordering gathers the main contributions towards the matrix main diagonal, off-diagonal tiles can be further compressed leading to superior compression capabilities over the original cable-by-cable sorting and other matrix re-arrangements. Using frequency matrices up to 48 Hz (i.e., 220 matrices with a total size of 712GB), we then perform time-domain MDD with 30 iterations of LSQR. The resulting deconvolved wavefield is shown in Figure 2 alongside the cross-correlation (i.e., adjoint) wavefield and the directly modelled reflectivity. Although the data has been compressed by a factor of 13.33 from its original size, the performance of MDD is not affected with most of the free-surface multiples present in the adjoint solution clearly

being removed.

To demonstrate the impact of the accuracy threshold, we present an additional MDD result using a much looser accuracy tolerance ($acc = 7e-4$) for the TLR compression of the frequency matrices (Fig.2c). Comparing this with Fig.2b, it becomes evident that reducing the accuracy to achieve higher compression levels introduces undesirable noise in the solution. We summarize the effect of tile size *nb* and accuracy *acc* in Fig. 3, where two opposing trends emerge. As we loosen the accuracy threshold (from $1e-4$ to $7e-4$), we trade the quality of the final solution for increased compression. Three regions are therefore identified: green, orange, and red, representing accurate, satisfactory (with some additional noise), and unacceptably inaccurate solutions, respectively. As a rule of thumb, the level of accuracy required by the MDD process depends on the downstream application as well as the computational resources available to the user. We consider the green solutions to be accurate enough for subsequent quantitative analysis, such as seismic inversion, while the yellow solutions may still be suitable for qualitative analysis, like seismic interpretation.

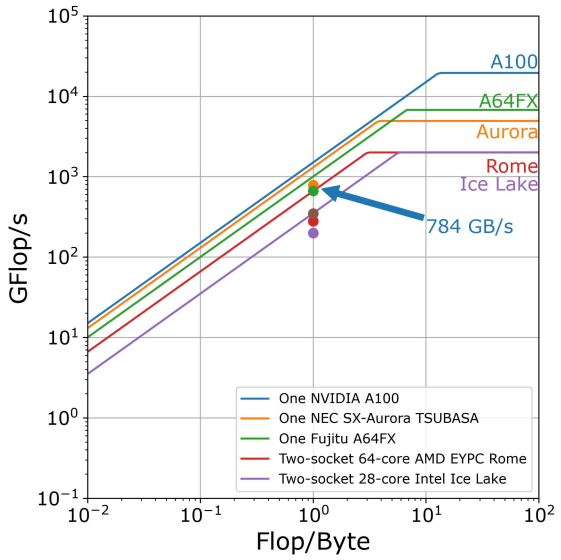

Figure 4: Roofline Performance Model for the TLR-MVM.

To conclude, we conduct a performance campaign of TLR-MVM on all mainstream architectures in the market, as MDC operators are the most time consuming part in the MDD application. We choose nb and error threshold that best fit the hardware specifications. Figure 4 presents the roofline performance models of various contemporary vendor offerings required to host our real seismic processing workload in memory. It includes single devices from NVIDIA, NEC, and Fujitsu, as well as two-socket x86 solutions from AMD and Intel. The results show that NVIDIA A100 is able to achieve around 800 GB/s bandwidth, which is the fastest among other 5 architectures.

CONCLUSION

We present the first application of time-domain MDD on a large-scale, geologically realistic 3D synthetic seismic dataset. To begin with, the data sparsity nature of the frequency-domain representation of the down-going wavefield is assessed: empirically, we observe that by applying a distance-aware reordering method (e.g., Hilbert sorting) to the rows and columns of the matrices to be compressed is critical to achieve decent compression whilst retaining the required accuracy. Our numerical results confirm that time-domain MDD is robust to the small numerical errors introduced by the tile low-rank compression algorithm. We reach around 800 GB/s memory bandwidth on a single TLR-MVM using a single A100 GPU. We benchmark the whole MDD application and finish the whole pipeline within 2 minutes using 4 A100 GPU (including I/O).

In future work, we plan to extend the algorithm to multi virtual source. The corresponding matrix vector multiplication will be changed into matrix matrix multiplication.

ACKNOWLEDGMENTS

The authors thank King Abdullah University of Science and Technology (KAUST) for funding his work. For computer time, this research used the resources of the Supercomputing Laboratory at KAUST in Thuwal, Saudi Arabia. The code used to perform TLR-MDC and TLR-MVM (Figure **??**) can be accessed at Github TLR-MDC and Github TLR-MVM.

REFERENCES

Aminzadeh, F., J. Brac, and T. Kunz, 1997, 3d salt and overthrust models: Presented at the SEG/EAGE Modeling Series, SEG Book Series Tulsa, Oklahoma.

Amundsen, L., 2001, Elimination of Free-surface Related Multiples Without Need of a Source Wavelet: Geophysics, **66**, 327–341.

Boiero, D., and C. Bagaini, 2020, Up-down deconvolution in complex geological scenarios: Online EAGE Conference and Exhibition, Extended Abstracts. (doi: 10.3997/2214-4609.2020611021).

Hong, Y., H. Ltaief, M. Ravasi, L. Gatineau, and D. Keyes, 2021, Accelerating seismic redatuming using tile low-rank approximations on nec sx-aurora tsubasa: Supercomputing Frontiers and Innovations, **8**, no. 2. (doi: 10.14529/jsfi210201).

Jakubowicz, H., 1998, Wave equation prediction and suppression of interbed multiples: Society of Exploration Geophysicists. (doi: 10.1190/1.1820204).

Lopez, G. A., and D. Verschuur, 2015, 3d focal closed-loop SRME for shallow water: Geophysical Journal International, **203**, 792–813. (doi: 10.1190/segam2015-5921009.1).

Ltaief, H., J. Cranney, D. Gratadour, Y. Hong, F. Ferreira, , and D. Keyes, 2021, Meeting the real-time challenges of ground-based telescopes using low-rank matrix computations: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 1–16. (doi: 10.1145/3458817.3476225).

Ravasi, M., Y. Hong, H. Ltaief, D. Keyes, and D. Vargas, 2022a, Large-scale Marchenko imaging with distance-aware matrix reordering, tile low-rank compression, and mixed-precision computations: Second International Meeting for Applied Geoscience & Energy, 2606–2610.

Ravasi, M., T. Selvan, and N. Luiken, 2022b, Stochastic multi-dimensional deconvolution: IEEE Transactions on Geoscience and Remote Sensing, **60**.

Ravasi, M., and I. Vasconcelos, 2021, An open-source framework for the implementation of large-scale integral operators with flexible, modern hpc solutions - enabling 3d marchenko imaging by least-squares inversion: Geophysics, **86**, WC177–WC194. (doi: 10.1190/geo2020-0796.1).

Ravasi, M., I. Vasconcelos, A. Curtis, and A. Kritski, 2015, Multi-dimensional free-surface multiple elimination and source deblending of Volve OBC data: 77th Conference and Exhibition, EAGE, Extended Abstracts. (doi: 10.3997/2214-4609.201413355).

van der Neut, J., and F. Herrmann, 2013, Interferometric redatuming by sparse inversion: Geophysical Journal International, **192**, 666–670. (doi: 10.1093/gji/ggs052).

van Groenestijn, G. J., and D. J. Verschuur, 2009, Estimating primaries by sparse inversion and application to near-offset data reconstruction: Geophysics, **74**, no. 3, 1MJ–Z54. (doi: 10.1190/1.3111115).

Vargas, D., I. Vasconcelos, M. Ravasi, and N. Luiken, 2021, Time-domain multidimensional deconvolution: A physically reliable and stable preconditioned implementation: Remote Sensing, **13**.

Vasconcelos, I., M. Ravasi, A. Kritski, J. van der Neut, and T. Cui, 2017, Local, reservoir-only reflection and transmission responses by target-enclosing extended imaging: SEG Technical Program Expanded Abstracts, 5289–5293. (doi: 10.1190/segam2017-17730961.1).

Verschuur, D. J., 1992, Surface-related multiple elimination in terms of Huygens sources: Journal of Seismic Exploration, **1**, 49–59.

Wapenaar, K., J. Thorbecke, J. van der Neut, F. Broggini, E. Slob, and R. Snieder, 2014, Marchenko imaging: Geophysics, **79**, no. 3, WA39–WA57. (doi: 10.1190/geo2013-0302.1).

Wapenaar, K., J. van der Neut, E. Ruigrok, D. Draganov, J. Hunziker, E. Slob, J. Thorbecke, and R. Snieder, 2011, Seismic interferometry by crosscorrelation and by multidimensional deconvolution: A systematic comparison: Geophysical Journal International, **185**, 1335–1364. (doi: 10.1111/j.1365-246X.2011.05007.x).