Responsibly Reckless Matrix Algorithms for HPC Scientific Applications

Hatem Ltaief, Marc G Genton, Damien Gratadour, David E Keyes, Matteo Ravasi

Queries to the Author

Q1. Please provide the postal code in the affiliation of the author Damien Gratadour.
Q2. Please provide the year of the refs. [19] and [20].
Q3. Please update Refs. [21] and [24].
Q4. Please provide page range for Refs. [23] and [27].
Q5. Please provide DOIs for all the references.

Responsibly Reckless Matrix Algorithms for HPC Scientific Applications

Hatem Ltaief and Marc G. Genton, *King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia*

Q1 Damien Gratadour, *Observatoire de Paris, Paris, France*

David E. Keyes and Matteo Ravasi, *King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia*

High-performance computing (HPC) achieved an astonishing three orders of magnitude performance improvement per decade for three decades, thanks to hardware technology scaling resulting in an exponential improvement in the rate of floating point executions, though slowing in the most recent. Captured in the Top500 list, this hardware evolution cascaded through the software stack, triggering changes at all levels, including the redesign of numerical linear algebra libraries. HPC simulations on massively parallel systems are often driven by matrix computations, whose rate of execution depends on their floating point precision. Referred to by Jack Dongarra, the 2021 ACM A.M. Turing Award Laureate, as "responsibly reckless" matrix algorithms, we highlight the implications of mixed-precision (MP) computations for HPC applications. Introduced 75 years ago, long before the advent of HPC architectures, MP numerical methods turn out to be paramount for increasing the throughput of traditional and artificial intelligence (AI) workloads beyond riding the wave of the hardware alone. Reducing precision comes at the price of trading away some accuracy for performance (reckless behavior) but in noncritical segments of the workflow (responsible behavior) so that the accuracy requirements of the application can still be satisfied. They offer a valuable performance/accuracy knob and, just as they are in AI, they are now indispensable in the pursuit of knowledge and discovery in simulations. In particular, we illustrate the MP impact on three representative HPC applications related to seismic imaging, climate/environment geospatial predictions, and computational astronomy.

Discoveries in computational science and engineering are often the result of multidisciplinary research that synergistically combines efforts from experts in hardware architecture, numerical libraries, system software, and domain science. The incentives for hardware and domain science experts are often orthogonal: extracting the expected performance for the former, and getting high-throughput accurate scientific outcomes for the latter. The developers of software libraries usually function as a bridge between the two communities, which ultimately requires all actors to move outside of their comfort zones. This may translate into taking "shortcuts" to achieve the desired outcomes, but these must still be rigorously verified.

Mixed-precision (MP) matrix algorithms are numerical methods that employ low/high precisions for storage and/or computations in noncritical/critical sections of the algorithm, respectively. Introduced 75 years ago,[1] then revisited in the context of solving a system of linear equations[2] and eigensolvers,[3] these MP methods reduce time-to-solution while recovering the lost numerical

1521-9615 © 2022 IEEE
Digital Object Identifier 10.1109/MCSE.2022.3215477

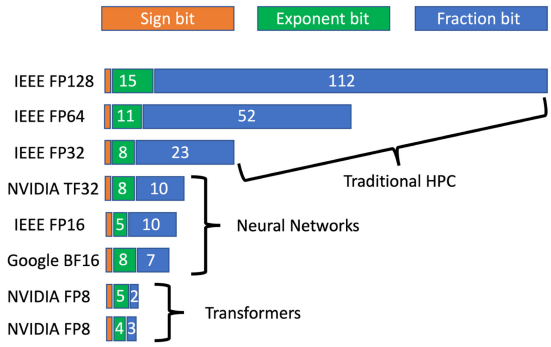

FIGURE 1. Overview of FP representations.

accuracy via an iterative refinement procedure. These early works on MP algorithms were substantial in providing rigorous error analysis and numerical stability studies. It is interesting that the development of MP algorithms occurred well before the emergence of high-performance computing (HPC) systems featuring dedicated hardware units for MP computations. Indeed, HPC history shows that hardware technologies are typically deployed first before HPC applications receive enough attention to effectively run on them. Putting the cart before the horse is not always smooth; hence recent efforts to establish a roadmap for software/hardware co-design.[4]

Although MP algorithms have existed for decades, their adoption into mainstream applications has gained momentum only lately, driven by the ever-increasing industry-led AI market. Figure 1 shows new floating-point (FP) representations designed by vendors specifically for AI workloads in addition to the existing IEEE 754 formats. The number of FP representations indicates the opportunity for the HPC community to embrace more flexibility in the software stack. Chip manufacturers have deployed GPU accelerators with special hardware support for fast MP arithmetic, attaining up to 30X performance speedup compared to FP64 arithmetic (see Table 1 in https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf[5]). This unprecedented performance improvement has initiated a forced march toward integrating MP matrix algorithms into traditional HPC scientific applications. This trend has been further accelerated by considerations related to energy consumption due to expensive data movement. Unfortunately, this may sometimes lead to the reckless usage of MP algorithms, without proper numerical validation, especially in situations where multidisciplinary collaboration may not be fostered. Recent work[6] provides mathematical insight and error analyses that will eventually engender a wider adoption from domain scientists.

Reckless but responsible MP matrix algorithms are what is needed to successfully bring together all actors in the exciting upcoming HPC era in which approximate computing will play a major role for scaling up scientific computing. The Innovative Computing Laboratory at the University of Tennessee, Knoxville, USA, led by Jack Dongarra, is a pioneer research institution in the development of numerical libraries for massively parallel systems, including MP matrix algorithms,[7] as implemented in *LAPACK*,[8] *ScaLAPACK*,[9] *DPLASMA*,[10] *PLASMA/MAGMA*,[11] and more recently *SLATE*.[12] The development of these libraries is more the result of a marathon than a sprint, with contributions from the community, perpetual algorithmic innovations, and performance portability across hardware vendors and generations.

JACK IN THE XBOX

The gaming industry represents one of the main markets for GPU hardware accelerators. While the main duty of these devices is to render graphics at high resolutions, GPUs have been used from early on as a commodity for performing traditional computational simulations. They were initially provisioned with low support for FP64 (in favor of FP32) with error correcting code memory disabled. This did not prevent a team of linear algebra experts, led by Jack Dongarra, to exploit the performance of FP32 arithmetic in obtaining FP64 accuracy. Early investigations started with using Intel's Pentium processors, AMD's Opteron architectures, and the IBM's Cell Broad Engine processor [12], which powered the famous Xbox and PlayStation gaming consoles. Their MP approach used an iterative refinement procedure to recover the precision loss while solving systems of linear equations, but keeping the bulk of the computation in FP32 to maximize the GPU's throughput. This work was among the first to democratize GPUs for traditional HPC workloads. But the road ahead was still bumpy for leveraging MP algorithms into applications.

"ISN'T IT DONE YET?"

As often asked by Jack Dongarra—but it takes a village to raise a child, and it takes the right ecosystem for sustainable research to develop. For MP computations, it was important for the hardware/software ecosystem to reach a certain level of maturity before MP numerical methods could become mainstream. MP algorithms realize the greatest advantage for their extra complexity when the execution rates widen between successive

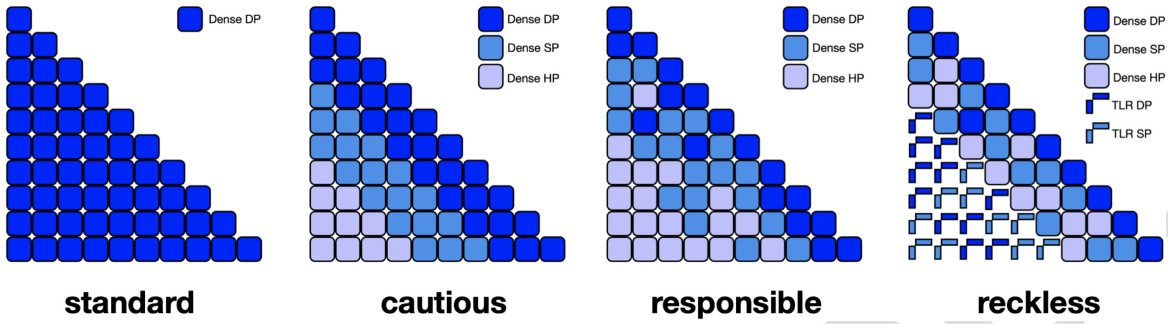

FIGURE 2. Tile algorithms adaptation for MP: An example with the Cholesky factorization.

precisions (see Table 1 in https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf[5]). AI algorithms revolutionized the way we do simulations, casting most of the underlying operations into low-precision matrix–matrix multiplications (GEMM), using FP8/FP16/BF16 FP representations. Chip manufacturers supported this trend by provisioning dedicated hardware units, which translated into unprecedented performance improvement. This is where research on MP algorithms began to flourish and impact the sustained performance of traditional HPC applications.

At the same time, the linear algebra community had to address the computational challenges brought to the fore by the manycore era with massively parallel hardware architectures. A profound redesign of the dense block algorithms available in the legacy *LAPACK* codes occurred to expose more fine-grained parallelism and mitigate the artifactual synchronization barriers. The resulting "tile algorithms" operate on the dense matrix split into small tiles using a task-based programming model. These computational tasks and their data dependencies constitute the vertices and edges of a directed acyclic graph, which is used to characterize the dataflow of the algorithm. A dynamic runtime system is then employed to orchestrate the task scheduling and execute them in an asynchronous fashion while ensuring data dependencies are not violated. This fine-grained task scheduling permits maximizing hardware occupancy and strong scaling in the presence of a large number of resources. As implemented in *PLASMA*[11]/*DPLASMA*[10]/*SLATE*,[12] these tile algorithms became standard for solving dense linear algebra problems.

To integrate MP into tile algorithms, the key idea is to extend the mathematical philosophy of iterative refinement[1,2,3] and to leverage the fine-grained task scheduling for performance, while reducing the overhead of data movement by transferring fewer bytes. This procedure consists in using low FP arithmetic for the bulk of the computation (i.e., matrix factorization) and refining/correcting the residual/solution using a higher FP arithmetic. A recent study[13] demonstrates the robustness of the procedure using three IEEE 754 FP16/FP32/FP64 precisions, even in presence of matrices with high condition numbers. This may make the MP with iterative refinement procedure agnostic to the application[7] in practice. These mathematical foundations are needed to design MP algorithms responsibly reckless. The MP with iterative refinement procedure does however require storing the whole matrix in the different precisions, which incurs a significant cost in terms of memory footprint.

NEW MP PERSPECTIVE

A new family of MP tile algorithms has emerged that exploits the data sparsity structure of the matrix, e.g., arising from Schur complements within discretizations of elliptic and parabolic PDEs, radial basis functions from unstructured meshing, and covariances in statistics.[14,15] In particular, the latter class of problems appears when modeling major scientific applications related to seismic imaging, climate/environmental geospatial predictions, and computational astronomy. This involves a symmetric matrix that represents correlations between data values of the physical phenomenon of interest. With proper ordering, strong correlations are typically located around the matrix diagonal and they start fading out as they are further away from the diagonal. Intuitively, when weak elements are combined with strong ones, their less significant bits fall off the edge of the result. They can therefore be approximated using lower FP arithmetic. Figure 2 shows a *standard* tile symmetric matrix (only the lower part is represented) stored in FP64 double precision. By leveraging the insights from the application, one can take advantage of the gradient pattern of the correlations and define band regions for applying corresponding FP arithmetic. This approach is referred to as *cautious* in Figure 2. This

approach is suboptimal in that the band tiles exhibiting strong correlations may be far from the diagonal when solving 3-D problems, hence requiring broad bands. A more adequate approach is to prescribe the FP precision on a tile-by-tile basis, referred to as *responsible* in Figure 2. This tile-centric approach relies on comparing the Frobenius norm of a tile against the Frobenius norm of the overall matrix, as explained in Higham and Mary's work.[6] Depending on the ratio of norms and the required numerical accuracy by the application, a tile-centric decision is made on the FP precision before the matrix operations proceed. To further overcome challenges of extreme scale, tile low-rank (TLR) matrix approximations can be applied in addition to MP techniques. Each tile can be independently compressed up to a prescribed accuracy threshold to reduce the overall memory footprint. In fact, TLR matrix approximations and MP can be combined, resulting in a representation of the matrix that is rather challenging to manipulate. We refer to this novel approach as *reckless* in Figure 2. The "recklessness" is not from the lack of mathematical foundations, but from the "wild west" of marshaling matrix operations on a tile data structure whose tiles can be stored in either dense or low-rank formats in any number of precisions. This is where tools based on runtime systems like *PaRSEC*[16] are key. They increase the user-productivity for code development and deal transparently with data movement of this *reckless* MP+TLR algorithms, beyond their traditional role of monitoring task scheduling.

Effectively and safely employing MP+TLR algorithms requires multidisciplinary expertise, engaging hardware architects, linear algebra algorithm developers, software engineers, and domain scientists. We present three scientific applications to highlight the matrix operations required to leverage these *cautious*, *responsible*, and *reckless* variants of MP algorithms:

1) imaging the Earth's subsurface using seismic redatuming, which requires fast matrix-vector multiplication;
2) modeling climate/environment with geostatistics; and
3) understanding the origin of life and the Universe, which both employ the Cholesky factorization as a preliminary step toward solving a linear system of equations and evaluating a matrix determinant.

We thus highlight opportunities and implications for applications that model physical phenomena from the Earth's subsurface, the Earth's atmosphere, and beyond the Earth.

IMAGING THE EARTH's SUBSURFACE

As early as the ancient Greeks, the curiosity of humankind for what lies beneath our feet has led to the development of the field of geophysics, a discipline that studies the Earth's structure and behavior. Like a camera, reflection seismology allows geophysicists to image the Earth's subsurface at meter-scale resolution.

The process of acquiring, processing, and imaging seismic data for large fields can take up to several weeks to months. Moreover, while a seismic acquisition campaign back in the 90s may have created data on the order of a few gigabytes, acquisition surveys are now producing terabytes (or even petabytes) of data. This creates computational challenges that make seismology of great interest to the HPC community. It is no surprise that some of the fastest supercomputers in the Top500 list such as Total's Pangea and ENI's HPC4 supercomputers have been specifically designed to tackle such challenges.[17]

As far as data storage and manipulation are concerned, the oil industry has been very conservative throughout the years. The main storage format, SEG-Y,[18] was developed in the mid 70s, a different *geological* era when we look at it through the lens of scientific computing; however, one specification of the SEG-Y format still dictates how seismic data are stored and processed to this day: each recorded seismic amplitude is in fact delivered to us in a 32-bit (FP32) format. When it comes to visualization and interpretation, seismic volumes are usually optimized for performance (e.g., DUG Insight User Manual–Optimizing Volumes for Performance[19]), meaning that copies of the same data are created in lower precision—usually 16-bit or even 8-bit. While not space-friendly, today's low cost of storing data motivates such an approach to ultimately improve user experience. But for quantitative analysis, geophysicists seem to be more conservative. Quoting field expert Matt Hall, *"...for seismic analysis, 8-bit data is probably not precise enough. Opinions vary, but I usually keep my 32-bit volume on disk, but make all derivative volumes and attribute volumes 16-bit. I think 65,536 values is enough, and because of noise and other uncertainties in the data, any precision beyond that is spurious."*[20]

Recent research in the area of seismic processing suggests that Matt may have been too conservative in his statement. Various seismic processing algorithms have been shown to be robust against storing the frequency domain representation of the seismic datasets in low-precision, alongside performing algebraic compression in the form of TLR approximations.[21] This is

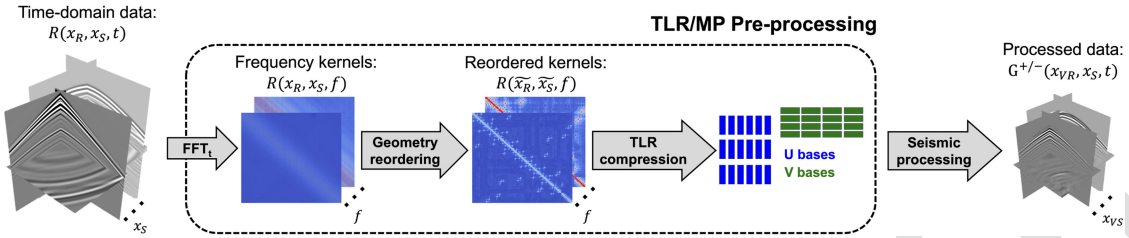

FIGURE 3. Schematic diagram of a responsibly reckless seismic processing workflow, where the input dataset is preprocessed into a set of low precision $\mathbf{U}$ and $\mathbf{V}$ bases to enable TLR based matrix-vector multiplications.

exactly the domain where such algorithms operate,[22] where a multidimensional convolution modeling operator that involves the evaluation of an extremely expensive complex-valued, batched matrix-vector multiplication operation is repeatedly applied. One can, therefore, afford to spend a certain amount of time upfront to apply a series of preprocessing steps such that the data is optimally arranged for subsequent efficient computations; Hong et al.'s work[21,23,24] show that by using TLR compression, matrix rearrangement based on Hilbert sorting, and low-precision storage as a way to reduce the sheer size of seismic data could lead to an extraordinary reduction in both computational resources and time. Figure 3 illustrates the complex seismic workflow using TLR-MVM, which eventually calls an MP procedure for further optimizing execution and data movement.

Using the synthetic seismic dataset created in Ravasi and Vasconcelos's work,[22] Hong et al.[24] compared the time-to-solution of TLR-MVM with and without Hilbert sorting and using different levels of precisions on a variety of architectures. In general, a speedup is observed in all cases compared to the standard dense MVM operation with FP32 precision. In particular, our TLR-MVM outperforms dense MVM by more than 6X on NVIDIA A100 GPU using a variety of precisions in compute and storage (i.e., INT8/FP16/FP32). Moreover, considering the entire process of redatuming (or virtually moving) seismic data from the Earth's surface[25] to an area of about 1.3 km^2 at a depth of 650 m below a complex overburden, the time breakdown in Figure 4(b) reveals that compression is just a tiny fraction of the overall cost of such a seismic processing step. And by continuing to develop this MP algorithm, the TLR-MVM time profile can be further reduced, while minimally impacting the overall quality of the results.

Was anything lost in terms of accuracy? Yes, as there is no free lunch in MP computations. Nevertheless, Figure 5 shows that the solution obtained at one subsurface point using a TLR compressed version of the seismic data and then mixing INT8/FP16/FP32 precisions in compute and storage is very similar to that obtained from its dense, FP32 counterpart. The signal-to-noise ratio is only 0.4-dB smaller in the low-precision case, compared to the "ground truth," although the data size has been compressed by a factor of over 100X! This reckless but responsible seismic processing workflow indicates that the geophysicist community may have been overcomputing and storing for years. Similar MP opportunities may be lying ahead for the seismic field at large.

MODELING CLIMATE/ENVIRONMENT WITH GEOSTATISTICS

Geostatistics models and predicts quantities of interest from data distributed in space-time based on statistical assumptions. It can be seen as complementary

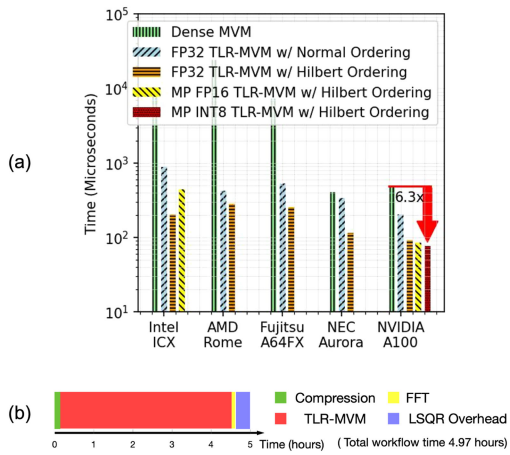

FIGURE 4. Performance results in terms of time-to-solution (a) for a single MVM using dense and TLR compressed matrices with different precisions, and (b) for an entire seismic redatuming workflow using the best choice of TLR compressed matrix kernel.

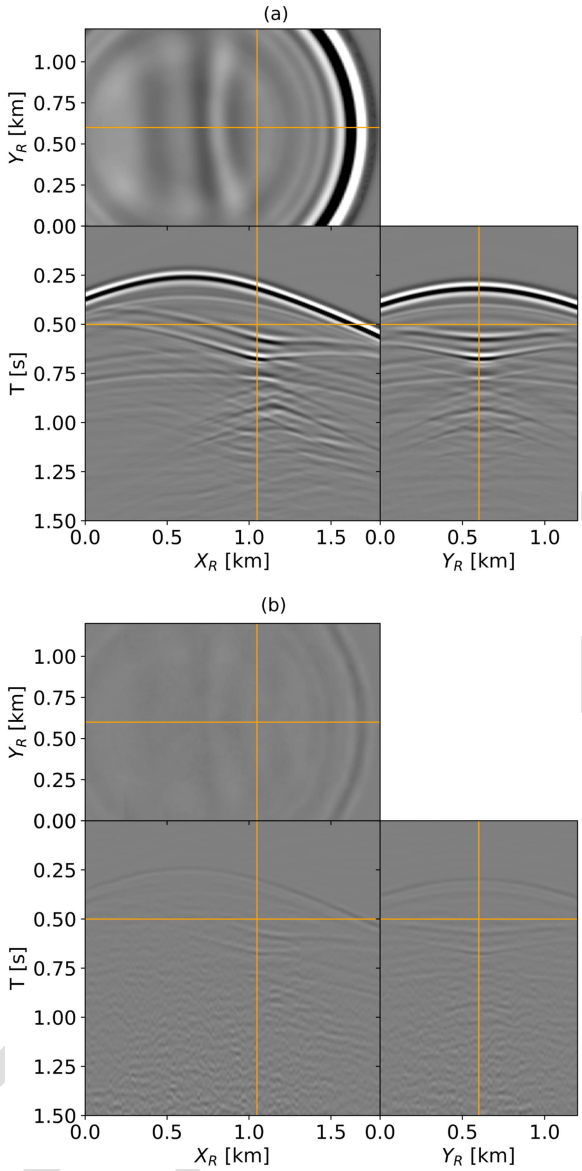

FIGURE 5. (a) Seismic redatuming wavefield estimate using dense, high-precision (FP32) matrix kernels. (b) Error introduced when switching to TLR compressed bases stored in INT8 precision.

to modeling approaches based on first principles rooted in conservation laws and physics-based models commonly expressed by PDEs. Climate and environmental applications, e.g., soil moisture variables recorded at the topsoil of the Mississippi River basin in Figure 6, are among the main workloads keeping supercomputers busy worldwide and are intended for exascale computers, thus even minor enhancements for production applications may provide significant rewards.

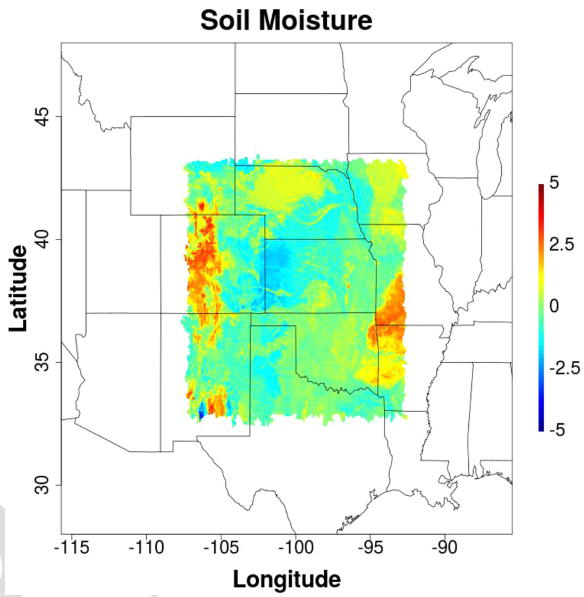

FIGURE 6. Soil moisture residuals at the topsoil of the Mississippi River basin.

A key ingredient of geostatistics is the covariance matrix, often based on an underlying spatial covariance function of the Matérn class, which appears in the likelihood function of Gaussian random fields, in the optimal spatial interpolation coined "Kriging" and its uncertainty quantification, and in the simulation of realizations from Gaussian random fields. This dense covariance matrix has a symmetric and positive-definite form, and its algebraic dimension equals the number of spatially distributed data values times the number of time steps. In the aforementioned tasks, two fundamental operations on the covariance matrix are the application of its inverse and the computation of its determinant. These operations can all be obtained through the celebrated Cholesky factorization and triangular solution, but are characterized by cubic/square complexity in the number of data values in flops/memory, respectively, and hence become unfeasible for large-scale problems. Indeed, a covariance matrix for 1M spatial data values would require 4-TB memory in (symmetric) DP format and on the order of 10^{18} flops to factor.

ExaGeoStat[26] is a software package built to provide user-controlled approximations to extreme-scale geostatistical problems by introducing innovative algorithmic, architectural, and programming model features. A MP tile Cholesky algorithm is introduced to speed up the factorization in the key geostatistical tasks. It is then deployed on large-scale heterogeneous systems with the help of *PaRSEC* dynamic runtime system.[16]

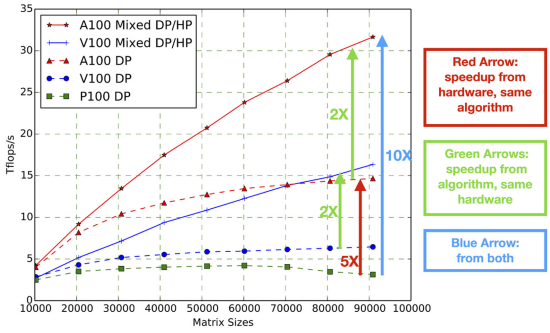

FIGURE 7. Gaussian maximum likelihood estimation performance breakdown across GPU generations and precision arithmetic.

With a suitable ordering, the algorithm works with double-precision arithmetic on tiles neighboring the main diagonal, while operating with single-precision or lower arithmetic for tiles far enough, leading to a three-precision FP16/FP32/FP64 approximation algorithm for the Cholesky factorization.[15] Referred to as *cautious* algorithms in Figure 2, *ExaGeoStat* leverages the inherent band data sparsity structure of the covariance matrix to exploit all FP representations accordingly. Figure 7 demonstrates the impact of MP tile algorithms across various NVIDIA GPU generations. The dashed curves show the progress that can be obtained by riding the hardware alone, through three generations of NVIDIA GPUs, i.e., Pascal, Volta, and Ampere, respectively. The red arrow shows a fivefold improvement from Pascal to Ampere GPUs. The solid curves show the performance obtained on the latter two architectures supporting a mixed DP and HP covariance matrix, with twofold improvements as shown by the green arrows. The combined improvement coming from hardware and algorithm, shown by the blue arrow, is tenfold. Numerical accuracy assessment[15] shows that with a proper band structure capturing the regions of strong/medium/weak correlations, *ExaGeoStat* is able to compute the relevant statistical parameters as if all computations were performed in FP64.

Finally, one can also leverage TLR approximations in addition to MP tile algorithms to further reduce the memory footprint and time-to-solution, leading to a *reckless* but *responsible* algorithm that can still ensure adequate accuracy.[6] With the help of *PaRSEC* dynamic runtime system,[16] the Cholesky factorization relies on a hybrid data distribution that mitigates the load imbalance between tasks next to and far from the diagonal with high/low algorithmic complexity, respectively. *PaRSEC* then marshals the data movement of tiles stored in dense and compressed formats. It also performs precision conversion on-the-fly and uses advanced look-ahead techniques to shorten the critical path, beyond its original duty of orchestrating task scheduling. This MP+TLR combination permits to take the best of the two worlds: 1) MP tile-dense algorithms applied on computational tasks around the critical path to shorten it and 2) TLR algorithms applied on the remaining computational tasks to reduce memory footprint and address big data problems. Preliminary results of the resulting responsibly reckless MP+TLR algorithms show that the Cholesky factorization may achieve an order of magnitude performance higher than if MP is applied alone on the original FP64 dense covariance problems. The numerical validation is even more robust with this algorithmic variant than the *cautious* version, especially when dealing with 3-D problems. Moving forward, this creates new opportunities to study parallel space-time likelihood optimization on large-scale systems,[27] which would be otherwise intractable.

CHASING THE ORIGIN OF LIFE AND THE UNIVERSE

The superior angular resolution provides exciting opportunities for astronomy, making possible major scientific breakthroughs by enabling better photometric and astrometric precision and better contrast. The 2020 Nobel Prize in Physics was, for instance, awarded to a group of astronomers "for the discovery of a supermassive compact object at the center of our galaxy," believed to be a giant black hole.[28] Within the observational arsenal, adaptive optics (AO) stood out as a game changer to enable this significant leap in our understanding of the Universe. High angular resolution is key to making detailed studies of both our Earth's neighborhood, distant reaches revealing the early Universe, and everything in between. Moreover, when coupled to high contrast techniques, getting sharper images allows astronomers to study exoplanets in extrasolar planetary systems, over a range of evolutionary stages in order to probe the initial conditions for planetary formation, the evolution of planetary systems over various time-frames and possibly the emergence of life.

While the largest ground-based telescopes will soon reach 40-m diameter and provide the angular resolution and collecting area required to detect the first stars and first galaxies as well as faint rocky exoplanets through direct imaging, they must be equipped with appropriate apparatus to overcome optical distortions induced by atmospheric turbulence. AO technologies, dating back to the late 1980s, were developed for this purpose and are now essential for the largest optical

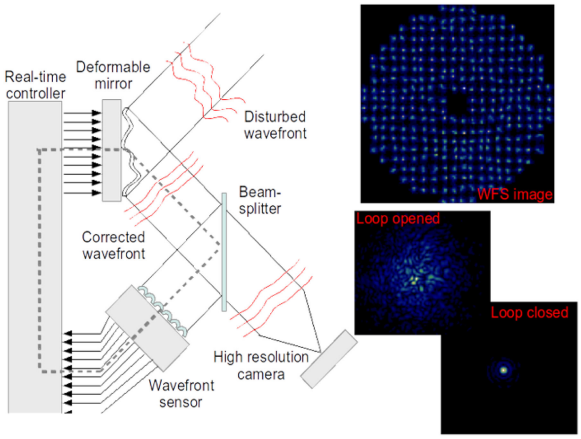

FIGURE 8. AO loop is composed of the DM, the WFS, and the RTC. A typical WFS image is shown on the upper right panel and images of a star in open and close loop operation models are shown below that.

telescopes. In its simplest form, an AO system is composed of a wavefront sensor (WFS) used to measure atmospheric distortions at a high frame rate, which are compensated with a deformable mirror (DM). The subsystem linking those components, responsible for interpreting wavefront measurements into actual commands to actuators of the DMs, is the real-time controller (RTC), as shown in Figure 8. It must operate at high speed (kHz rate) to catch up with the rapidly changing optical turbulence.

One of the limitations of classical AO is that the correction is only valid in a very small patch of sky, the size of which depends on the observing wavelength, from a few arcseconds in the visible to a few tens of arcseconds in the near infrared. Multiconjugate adaptive optics (MCAO) solves this problem by using a series of DMs to compensate the turbulence in volume, enabling AO correction over a wide field of view. MCAO uses several guide stars and associated WFSs to probe the light wave aberrations in several directions, and the RTC, using tomographic reconstruction, determines the best commands to apply to the DMs.[29]

The classical approaches to aberrations retrieval are based on linear models of the relationship between the sensor(s) data and the phase of the light wave, relying on careful modeling of the AO system error budget and solved through a linear approach, sometimes regularized with a given prior on the turbulence statistics. For instance, in present-day conventional AO systems, the RTC follows a well-defined linear control scheme: input measurements from sensors are multiplied by a control matrix to produce an output DM control command.

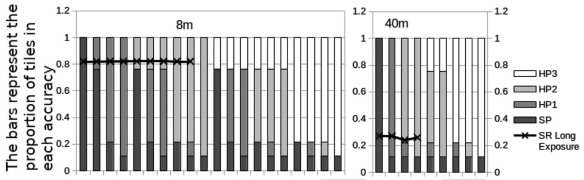

FIGURE 9. Proportion of the different GEMM precision and associated ToR accuracy on an 8-m telescope with 17K measurements (left) and a 40-m telescope with 50K measurements (right).

The computation of the tomographic reconstructor (ToR), which consists of the Cholesky factorization of the dense symmetric covariance matrix of WFS measurements followed by a backward and forward substitution, is at the core of operations for all tomographic AO instruments and must be updated regularly to take into account the evolution of the atmosphere's light-bending structure. The most time-consuming kernel during the factorization and the solve phase is the general matrix-matrix multiplication (GEMM), which makes the compute-bound algorithm run close to the system's sustained peak performance. Performance results have been reported on shared-memory systems equipped with hardware accelerators[30] as well as distributed-memory systems[31] using single precision FP arithmetic. The measurements used to generate the matrix operator come from 16-bit unsigned integer signals from WFS cameras, which has to be converted to 32-bit in order to perform FP computations of the covariance matrix. In addition, the dense covariance matrix, which may be of size as large as 100K, has a data-sparse structure, due to weak interactions between some of the measurements taken by the WFS. This is expected, since measurements taken by WFS subapertures physically located next to each other exhibit higher correlations, while as we move away from the matrix diagonal, weak interactions between measurements taken by remote WFS subapertures are expected.

The ToR computation has been tested for two different ranges of systems dimensions, including a state-of-the-art 8-m telescope instrument and a future 40-m telescope instrument up to a total of 17K and 50K measurements, respectively. Figure 9 shows the proportion of tiles operated on as single precision or the different variants of half-precision, as well as the performance of the computed MP ToR. This corresponds to the *cautious* MP variant of Figure 2 based on a band structure for determining the FP32 or FP16 arithmetic. The graph entries are sorted given the type and proportion of half-precision, and the ToR performance is expressed as

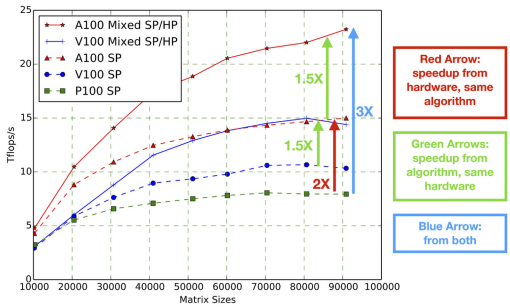

FIGURE 10. ToR performance breakdown across GPU generations and precision arithmetic.

Strehl ratio (SR): a measure of image quality as the ratio of the maximum value in the image over its theoretical maximum, 1 being the best achievable ratio. These SRs are obtained with the end-to-end simulation tool COMPASS[32] generating long exposure images from a full model of the system, including turbulence, telescope, and AO instrument. SR values are missing when the Cholesky factorization did not succeed (due to the loss of positiveness engendered by the low precision), but in a successful case, the ToR accuracy is almost equal to the full single precision approach, regardless of the considered instrument dimensioning.

Figure 10 tracks the same type of improvements across hardware generations and algorithmic improvements as in Figure 7, this time between SP and HP. The MP ToR computation on the newest Ampere GPU scores a threefold speedup compared to the reference SP ToR implementation on the previous Pascal GPU generation. The sustained performance achieved by using FP16/FP32 hardware support from the latest NVIDIA A100 GPU enables the computational astronomy community of ground-based telescopes to improve science insights during nightly on-sky observations. This result has implications for all instruments requiring real-time processing needed in AO.

Moreover, by using a tile-centric *responsible* MP variant, the ToR computation can capture the needed accuracy per tile[6] and ensure numerical robustness without requiring *a priori* knowledge on the data sparsity structure of the covariance matrix. Combining MP with TLR matrix approximations (i.e., *reckless* but *responsible* MP+TLR algorithms) is an interesting avenue to further satisfy real-time requirements and will be investigated in future work.

CONCLUSION

MP matrix algorithms have evolved significantly since their initial introduction 75 years ago. Recent work provides new analyses to take into account multiple lower precisions of FP arithmetic, which emerged with the advent of hardware technologies that support AI. Further challenges await with new non-IEEE FP representations (e.g., BF16 and FP8) fostered by chip manufacturers. Rigorous numerical validations are critical to ensure these reckless algorithms remain responsible. The new tile-centric MP approach for extreme-scale applications offers further opportunities to economize on storage and execution time while meeting user-specified accuracy requirements, which were often exceeded by default double precision. However, multidisciplinary expertise is required to navigate the software stack: the more reckless the MP algorithm, the more responsible the users must be.

Dense linear algebra is just one of the "seven dwarfs" identified by Phil Colella in a famous 2004 presentation[33] titled "Defining Software Requirements for Scientific Computing." It is time to review the amenability of the other six to exploiting the new precisions available in hardware in order to further stretch the capacity of HPC systems for extreme applications.

ACKNOWLEDGMENTS

The authors would like to thank Fujitsu/NVIDIA/NEC for the remote access to their respective systems and Intel/AMD for their hardware donations.

REFERENCES

1. L. Fox, H. D. Huskey, and J. H. Wilkinson, "Notes on the solution of algebraic linear simultaneous equations," *Quart. J. Mechanics Appl. Math.*, vol. 1, no. 1, pp. 149–173, 1948, doi: 10.1093/qjmam/1.1.149. Q5
2. C. B. Moler, "Iterative refinement in floating point," *J. ACM*, vol. 14, no. 2, pp. 316–321, Apr. 1967, doi: 10.1145/321386.321394.
3. J. J. Dongarra, C. B. Moler, and J. H. Wilkinson, "Improving the accuracy of computed eigenvalues and eigenvectors," *SIAM J. Numer. Anal.*, vol. 20, no. 1, pp. 23–45, 1983, doi: 10.1137/0720002.
4. J. Dongarra et al., "The International Exascale Software Project roadmap," *Int. J. High Perform. Comput. Appl.*, vol. 25, no. 1, pp. 3–60, Feb. 2011, doi: 10.1177/1094342010391989.
5. "NVIDIA A100 Tensor Core GPU Architecture V1.0," 2020. [Online]. Available: https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf Q2
6. N. J. Higham and T. Mary, "Mixed precision algorithms in numerical linear algebra," *Acta Numerica*, vol. 31, pp. 347–414, 2022, doi: 10.1017/S0962492922000022.

7. A. Abdelfattah et al., "A survey of numerical linear algebra methods utilizing mixed-precision arithmetic," *Int. J. High Perform. Comput. Appl.*, vol. 35, no. 4, pp. 344–369, 2021, doi: 10.1177/10943420211003313.
8. E. Anderson et al., *LAPACK Users' Guide*, vol. 9. Philadelphia, PA, USA: SIAM, 1999.
9. L. S. Blackford et al., *ScaLAPACK Users' Guide*. Philadelphia, PA, USA: SIAM, 1997.
10. G. Bosilca et al., "Flexible development of dense linear algebra algorithms on massively parallel architectures with DPLASMA," in *Proc. IEEE Int. Symp. Parallel Distrib. Process. Workshops Ph.D. Forum*, 2011, pp. 1432–1441.
11. E. Agullo et al., "Numerical linear algebra on emerging architectures: The PLASMA and MAGMA projects," *J. Phys.: Conf. Ser.*, vol. 180, no. 1, 2009, Art. no. 012037.
12. M. Gates, J. Kurzak, A. Charara, A. YarKhan, and J. Dongarra, "SLATE: Design of a modern distributed and accelerated linear algebra library," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, 2019, pp. 1–18, doi: 10.1145/3295500.3356223.
13. E. Carson and N. J. Higham, "Accelerating the solution of linear systems by iterative refinement in three precisions," *SIAM J. Sci. Comput.*, vol. 40, no. 2, pp. A817–A847, 2018, doi: 10.1137/17M1140819.
14. S. Abdulah, H. Ltaief, Y. Sun, M. G. Genton, and D. E. Keyes, "Geostatistical modeling and prediction using mixed precision tile cholesky factorization," in *Proc. IEEE 26th Int. Conf. High Perform. Comput., Data, Analytics*, 2019, pp. 152–162.
15. S. Abdulah et al., "Accelerating geostatistical modeling and prediction with mixed-precision computations: A high-productivity approach with PaRSEC," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 4, pp. 964–976, Apr. 2022.
16. G. Bosilca, A. Bouteiller, A. Danalis, M. Faverge, T. Hérault, and J. J. Dongarra, "PaRSEC: Exploiting heterogeneity to enhance scalability," *Comput. Sci. Eng.*, vol. 15, no. 6, pp. 36–45, 2013.
17. B. Felix, "Oil group total hopes new supercomputer will help it find oil faster and more cheaply, 2019. [Online]. Available: https://www.reuters.com/article/us-total-supercomputer-idUSKCN1TJ0FQ
18. K. Barry, D. Cavers, and C. Kneale, "Recommended standards for digital tape formats," *Geophysics*, vol. 40, no. 2, pp. 344–352, 1974.
19. DUG Insight User Manual—Optimising Volumes for Performance. [Online]. Available: https://help.dugeo.com/m/Insight/l/438665-optimising-volumes-for-performance
20. M. Hall, "B is for Bit depth." [Online]. Available: https://agilescientific.com/blog/2011/2/4/b-is-for-bit-depth.html
21. Y. Hong, H. Ltaief, M. Ravasi, D. Keyes, and D. Vargas, "Large-scale marchenko imaging with distance-aware matrix reordering, tile low-rank compression, and mixed-precision computations," *SEG Tech. Abstract*, 2022, accepted.
22. M. Ravasi and I. Vasconcelos, "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*, vol. 86, pp. WC177–WC194, 2021, doi: 10.1190/geo2020-0796.1.
23. Y. Hong, H. Ltaief, M. Ravasi, L. Gatineau, and D. Keyes, "Accelerating seismic redatuming using tile low-rank approximations on NEC SX-aurora TSUBASA," *Supercomputing Front. Innovations*, vol. 8, 2021, doi: 10.14529/jsfi210201. Q3
24. Y. Hong, H. Ltaief, M. Ravasi, and D. Keyes, "HPC seismic redatuming by inversion with algebraic compression and multiple precisions," *ACM Trans. Math. Softw.*, 2022, in preparation.
25. K. Wapenaar, J. Thorbecke, J. van der Neut, F. Broggini, E. Slob, and R. Snieder, "Marchenko imaging," *Geophysics*, vol. 79, pp. WA39–WA57, 2014, doi: 10.1190/geo2013-0302.1.
26. S. Abdulah, H. Ltaief, Y. Sun, M. G. Genton, and D. E. Keyes, "ExaGeoStat: A high performance unified software for geostatistics on manycore systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 12, pp. 2771–2784, Dec. 2018. Q4
27. M. L. Salvana, S. Abdulah, H. Ltaief, Y. Sun, M. Genton, and D. Keyes, "Parallel space-time likelihood optimization for air pollution prediction on large-scale systems," in *Proc. Platform Adv. Sci. Comput. Conf.*, 2022.
28. "Press release: The nobel prize in physics 2020," nobelprize.org, 2020. [Online]. Available: https://www.nobelprize.org/prizes/physics/2020/press-release/
29. F. Rigaut and B. Neichel, "Multiconjugate adaptive optics for astronomy," *Annu. Rev. Astron. Astrophys.*, vol. 56, no. 1, pp. 277–314, 2018, doi: 10.1146/annurev-astro-091916-055320.
30. H. Ltaief, D. Gratadour, A. Charara, and E. Gendron, "Adaptive optics simulation for the world's largest telescope on multicore architectures with multiple GPUs," in *Proc. Platform Adv. Sci. Comput. Conf.*, 2016, pp. 1–12, doi: 10.1145/2929908.2929920.
31. H. Ltaief et al., "Real-time massively distributed multi-object adaptive optics simulations for the european extremely large telescope," in *Proc. IEEE Int. Parallel Distrib. Process. Symp.*, May 2018, pp. 75–84.
32. F. Ferreira, D. Gratadour, A. Sevin, and N. Doucet, "Compass: An efficient GPU-based simulation software for adaptive optics systems," in *Proc. Int. Conf. High Perform. Comput. Simul.*, 2018, pp. 180–187.

33. P. Colella, "Defining software requirements for scientific computing. Slide of 2004 presentation included in David Patterson's 2005 talk," 2004. [Online]. Available: http://www.lanl.gov/orgs/hpc/salishan/salishan2005/davidpatterson.pdf

HATEM LTAIEF is the principal research scientist of the Extreme Computing Research Center, King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia. His research interests include parallel numerical algorithms, parallel programming models, and performance optimizations for multicore architectures and hardware accelerators. He is a Member of IEEE. Contact him at Hatem.Ltaief@kaust.edu.sa.

MARC G. GENTON is a distinguished professor of statistics with King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia. His research interests include statistical analysis, flexible modeling, prediction, and uncertainty quantification of spatio-temporal data, with applications in environmental and climate science, renewable energies, geophysics, and marine science. He is a fellow of the ASA, of the IMS, and the AAAS, and is an elected member of the ISI. Contact him at Marc.Genton@kaust.edu.sa.

DAMIEN GRATADOUR is an associate professor with Laboratoire d'Etudes Spatiales et d'Instrumentation en Astrophysique, Observatoire de Paris, Paris, France. His research interests include bridges astronomy with high-performance computing and artificial intelligence, applied to modeling, signal processing and instrumentation for large telescopes. Contact him at damien.gratadour@obspm.fr.

DAVID E. KEYES is the director of the Extreme Computing Research Center with King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia. He works with the interface between parallel computing and the numerical analysis of PDEs with a focus on scalable implicit solvers, such as the Newton–Krylov–Schwarz and the Additive Schwarz Preconditioned Inexact Newton methods, which he co-developed. He is a fellow of the SIAM, AMS, and AAAS. He is also a Member of IEEE. Contact him at David.Keyes@kaust.edu.sa.

MATTEO RAVASI is an assistant professor of geophysics with King Abdullah University of Science and Technology, Thuwal, 23955, Saudi Arabia. His research interests include geophysical inverse problems with applications to seismic acquisition and processing, imaging, quantitative interpretation, and time-lapse monitoring. He is also interested in the use of machine learning and high-performance computing and heavily involved in the development of open-source scientific software. Contact him at Matteo.Ravasi@kaust.edu.sa.