NAME **Bedros AFEYAN**

TITLE *Exploiting sparsity and machine learning in kinetic simulations of plasmas*

ABSTRACT We will show systematic improvement over Monte Carlo sampling of particle in cell codes due to smart sparse sampling techniques in nonlinear, kinetic simulations of plasma dynamics. The intelligent approach is learned and applied to nearby problems in parameter and resolution space readily, vastly speeding up families of simulations. Use of compression schemes will be discussed to carry the data from one simulation to a series of nearby simulations. Spectral and wavelet techniques will be applied.

NAME **Julien BRENNECK**

TITLE *Iteration for Contour-Based Nonlinear Eigensolvers*

ABSTRACT Contour integration techniques have become a popular choice for solving the linear and non-linear eigenvalue problems. They principally include the Sakurai-Sugiura methods, the Beyn's algorithm, the FEAST/NLFEAST algorithms and other rational filtering techniques. While these methods can result in effective 'black-box' approaches for solving linear eigenvalue problems, they still present several shortcomings for addressing nonlinear eigenvalue problems which are both mathematically and practically far more challenging. We introduce a new hybrid algorithm that advantageously combines the iterative nature of NLFEAST with the effectiveness of Beyn's approach to deal with general non-linearity. In doing so, this NLFEAST-Beyn hybrid algorithm can overcome current limitations of both algorithms taken separately. After presenting its derivation from both a Beyn's and NLFEAST's perspective, several numerical examples are discussed to demonstrate the efficiency of the new technique.

NAME **Jinhao CHEN**

TITLE *Sparse roundoff-error-free LU update*

ABSTRACT LU update is a key component of the simplex method, which is one of the most popular solvers for linear programming. Due to the growing need for extended precision or exact precision high, a dense roundoff-error-free (REF) LU update algorithm has been proposed. Different from conventional LU update algorithms that require rational arithmetic for exact computation, the dense REF LU update algorithm works entirely in integer arithmetic. However, most of the real world applications can be represented with sparse matrices, a REF LU update algorithm for sparse matrices is more critical. In this talk, theorems with rigorous mathematical proof and the proposed sparse REF LU update algorithm will be presented. Furthermore, preliminary results by the MATLAB implementation and further discussion on the data structure and expected time complexity will be included.

NAME **Chao CHEN**

TITLE *RCHOL: Randomized Cholesky Factorization for SDD Matrices*

ABSTRACT We present a randomized algorithm to construct a preconditioner for solving sparse symmetric diagonally-dominant (SDD) linear systems. It is a well-known fact that a Gaussian elimination/Cholesky step introduces a n-by-n dense fill-in block in the Schur complement if the eliminated row/column has n+1 nonzero entries (including a diagonal

entry). By contrast, we employ a randomized algorithm to sample only O(n) entries, which equal to the exact fill-in block in expectation. From a graph perspective, the n-by-n fill-in block corresponds to a clique, and the randomized strategy samples n-1 edges from the clique.

To optimize for practical performance, we first reorder the input sparse matrix to exploit sparsity. In particular, we apply a log2(p)-level nested dissection followed by the approximate minimum degree ordering at the leaf level, where p is the number of threads. This reordering naturally leads to a multifrontal-type parallel method.

Numerical results show that the randomized preconditioner has a small memory footprint, which is typically 3~4 times of the input sparse matrix arising from the discretization of a partial differential equation in three dimensions. With the same amount of memory, the preconditioner required much less iterations and wall-clock time than standard incomplete Cholesky with threshold dropping.

NAME **John CONROY**

TITLE *Two Truths in Spectral Clustering as Seen in Three Theorems and Three Applications*

ABSTRACT Spectral graph clustering, which clusters vertices of a graph by a spectral embedding and proba- bilistic mixture models, has many applications in the sciences and has a rich theoretical foundation. Stochastic block models provide a rich model of real-world graphs and one through which the theory of spectral clustering can be advanced. In this talk, we explore two commonly used spectral clus- tering methods, Adjacency Spectral Embedding (ASE) and Laplacian Spectral Embedding (LSE). These methods proceed by first performing a truncated eigenvalue value decomposition of a sparse symmetric matrix followed by K-means or, more generally, Gaussian mixture model estimation. Both ASE and LSE methods are known to be consistent for stochastic block models, i.e., both spectral decompositions give rise to clusters which can be used to converge to the parameters of an unknown stochastic block model from which the graph is sampled. Until recently, it was thought that both ASE and LSE, when applied to a graph would generate comparable graph clustering. However, in applying the methods to the human connectome, a graph model of the human brain, it was observed empirically that the LSE was more likely to find hemispheric partitions of the brain while ASE was more likely to discover the gray-white matter partitioning of the brain. After considerable effort, a theorem was stated and proven demonstrating that the ASE spectral clus- tering was more likely to discover the G/W (core-periphery) partition and LSE was more likely to discover the left hemisphere/right hemisphere affinity partition of the brain [4]. The two truths phenomenon has also been observed in computer networks and graphical models of relations of words in computational linguistic language models.

Beyond the human connectome these methods and the “two-truths” has been observed and ex- ploited in computer network security and computational linguistics. The network security structure is nicely seen in the Los Alamos National Labs network authentication dataset. These data include a “red-team” event, where malicious activity, i.e., attempts to “hack” into servers on the network is present. The analogues of L/R hemisphere (affinity) and G/W (core-periphery) partitions here are subnets and tightly connected areas of the network vs loosely connected. The red-team event was an attack on a set of highly used servers, and indeed a statistical classification method based on ASE outperforms one based on LSE [3].

In the context of computational linguistics the natural graph to consider is the co-occurrence of words with their “contexts,” e.g., the set of one or more words that

preceded a word as observed in a collection of text [1]. The spectral methods applied in the text application uncover topic models, clusters of words that are related for LSE for a set of documents and ASE will tend to find “key concepts” versus “less important” information (peripheral ideas).

[1] Yoshua Bengio, R´ejean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. J. Mach. Learn. Res., 3:1137–1155, March 2003.
[3] John M. Conroy. Classification of Red Team Authentication Events in an Enterprise Network, chapter Chapter 9, pages 179–194. World Scientific Press, 2019.
[4] Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Minh Tang, Avanti Athreya, Joshua Cape, and Eric Bridgeford. On a two-truths phenomenon in spectral graph clustering. Proceedings of the National Academy of Sciences, 116(13):5995–6000, 2019.

NAME **Jeffrey CORNELIS**

TITLE *Projected Newton method for the regularization of ill-posed linear inverse problems*

ABSTRACT Regularization is necessary to obtain a meaningful solution to an ill-conditioned linear least squares problem contaminated with measurement errors or noise. Tikhonov regularization is a well-established and well-studied technique and often gives satisfactory results, especially in combination with a suitable regularization matrix. For large-scale problems, however, it can be quite computationally expensive to compute the Tikhonov solution, certainly when the regularized problem has to be solved multiple times to calculate a suitable regularization parameter. In this talk we present an efficient algorithm that simultaneously computes the regularization parameter and corresponding Tikhonov regularized solution such that the discrepancy principle is satisfied. This problem can be formulated as a large-scale system of non-linear equations that can be solved using Newton’s method. We show that by projecting the problem on a low-dimensional (generalized) Krylov subspace and calculating the Newton direction for this much smaller problem, we get a good search direction for the full-dimensional problem that is much cheaper to compute than the true Newton direction. This (generalized) Krylov subspace-based approach is especially well-suited for linear inverse problems with a large and sparse system matrix. We discuss the extension of the algorithm to allow a general regularization term, such as total variation regularization, which is a popular technique used in image restoration. The behaviour and performance of our approach is illustrated by a number of numerical experiments.

NAME **Ieva DAUZICKAITE**

TITLE *Spectral estimates for saddle point matrices arising in weak constraint four-dimensional variational data assimilation*

ABSTRACT Data assimilation estimates the state of a dynamical system by combining a previous estimate of the state and observations of the system. It is used to determine initial conditions in numerical weather prediction, where the state can have 10^9 variables and 10^7 observations have to be assimilated [1, 2].

One approach to obtain the estimate of the state (called the analysis) is to solve a weighted least-squares problem, which requires minimizing a nonlinear cost function. We consider the weak constraint four-dimensional variational (4D-Var) data assimilation method, which uses a cost function that is formulated under the assumption that the numerical model is not perfect and penalizes the weighted discrepancy between the analysis and the observations, the analysis and the previous estimate, and the difference

between the analysis and the trajectory given by integrating the dynamical model. The minimum of the cost function is approximated using an inexact Gauss-Newton method [3], in which a series of linearized quadratic cost functions with a low resolution model are minimized [4].

We consider the large sparse symmetric linear systems of equations that arise in the minimisation of the quadratic cost functions. These can be written as saddle point systems with a 3x3 block structure but block eliminations can be performed to reduce them to saddle point systems with a 2x2 block structure, or further to symmetric positive denite systems. In this talk, we analyse how sensitive the spectra of these matrices are to the number of observations of the underlying dynamical system. We also obtain bounds on the eigenvalues of the matrices. Numerical experiments are used to confirm the theoretical analysis and bounds. More details can be found in [5].

References

[1] N. K. Nichols, "Mathematical concepts of data assimilation", in Data Assimilation. Making Sense of Observations (W. Lahoz, B. Khattatov, and R. Menard, eds.), pp. 13 - 39, Springer-Verlag Berlin Heidelberg, 2010.

[2] A. S. Lawless, "Variational data assimilation for very large environmental problems", in Large Scale Inverse Problems: Computational Methods and Applications in the Earth Sciences (M. Cullen, M. A. Freitag, S. Kindermann, and R. Scheichl, eds.), Radon Series on Computational and Applied Mathematics 13, pp. 55 - 90, De Gruyter, 2013.

[3] S. Gratton, A. S. Lawless, and N. K. Nichols, "Approximate Gauss-Newton methods for nonlinear least squares problems", SIAM Journal on Optimization, vol. 18, no. 1, pp. 106 - 132, 2007.

[4] P. Courtier, J.-N. Thépaut, and A. Hollingsworth, "A strategy for operational implementation of 4D-Var, using an incremental approach", Quarterly Journal of the Royal Meteorological Society, vol. 120, no. 519, pp. 1367 - 1387, 1994.

[5] I. Daužickaitė, A. S. Lawless, J. A. Scott, and P. J. van Leeuwen, "Spectral estimates for saddle point matrices arising in weak constraint four-dimensional variational data assimilation", Numerical Linear Algebra with Applications, vol. 27, no. 5, p. e2313, 2020.

NAME **Tim DAVIS**

TITLE *SuiteSparse:GraphBLAS: graph algorithms in the language of linear algebra*

ABSTRACT SuiteSparse:GraphBLAS is a full implementation of the GraphBLAS standard, which defines a set of sparse matrix operations on an extended algebra of semirings using an almost unlimited variety of operators and types. When applied to sparse adjacency matrices, these algebraic operations are equivalent to computations on graphs. GraphBLAS provides a powerful and expressive framework for creating graph algorithms based on the elegant mathematics of sparse matrix operations on a semiring. Key features and performance of the SuiteSparse implementation of GraphBLAS package are described. The implementation appears in Linux distros, and forms the basis of the RedisGraph module of Redis, a commercial graph database system. Graph algorithms written in GraphBLAS can rival the performance of highly-tuned specialized kernels, while being far simpler for the end user to write.

NAME **Andrei DUMITRASC**

TITLE *Deflation for the off-diagonal block in saddle-point systems*

ABSTRACT Deflation techniques are typically used to shift isolated clusters of small eigenvalues in order to obtain a tighter distribution and a smaller condition number. Such changes induce a positive effect in the convergence behavior of Krylov subspace methods, which are among the most popular iterative solvers for large sparse linear systems.
We develop a deflation strategy for saddle-point matrices by taking advantage of their underlying block structure. The vectors used for deflation come from an elliptic singular value decomposition relying on the generalized Golub-Kahan bidiagonalization process. The block targeted by deflation is the off-diagonal one since it features a problematic singular values distribution for certain applications. One example is the Stokes flow in elongated channels, where the off-diagonal block has several small, isolated singular values, depending on the length of the channel.
Applying deflation to specific parts of the saddle-point system is important when using solvers such as CRAIG, which operates on individual blocks rather than the whole system. The theory is developed by extending the existing framework for deflating square matrices before applying a Krylov subspace method like MINRES. Numerical experiments confirm the merits of our strategy and lead to interesting questions about using approximate vectors for deflation.

NAME **Azzam HAIDAR**

TITLE *How NVIDIA Tensor Cores can Help HPC Scientific Application Unleash the Power of GPUs using Mixed Precision Solvers*

ABSTRACT Reduced precision computation such as FP16, TF32 (NVIDIA TensorFloat-32) and BF16 have enjoyed an order of magnitude growth in computational throughput over the FP64 with the advent of Tensor Cores on GPUs.
In this talk we cover schemes to leverage these reduced and mixed-precision tensor cores to accelerate the most commonly used numerical methods, the solution of linear systems of equations (Ax=b) without sacrificing any accuracy compared to the full fp64 solvers for both dense and sparse systems. We achieved above 3X performance increase and 5x better energy efficiency versus the standard FP64 implementation while providing a solution to the FP64 accuracy.

NAME **Roman IAKYMCHUK**

TITLE *Conjugate Gradient Solvers with Accuracy and Reproducibility Guarantees in Hybrid Parallel Environments*

ABSTRACT On Krylov subspace methods such as the Conjugate Gradient (CG), the number of iterations until convergence may increase due to the loss of computation accuracy caused by rounding errors in floating-point computations. Besides, as the order of operations is non-deterministic on parallel computations, the result and the behavior of the convergence may be non-identical in different environments, even for the same input. Hence, we propose three algorithmic solutions to guarantee accurate and reproducible computations in CG:
- the first is based on the Ozaki scheme, which is the error-free transformation for dot-product that can ensure the correct-rounding. One of the benefits of the method is that

it can be built upon vendor-provided linear algebra libraries such as Intel Math Kernel Library and NVIDIA cuBLAS/ cuSparse, reducing the development cost;
- the second solution originates from the ExBLAS approach, which efficiently combines floating-point expansions and long accumulator, and preserves every bit of result until the final rounding;
- the third is a customized version of the second that relies upon floating-point expansions and, hence, expands the intermediate precision.
Instead of converting the entire solver, we identify those parts that violate reproducibility/ non-associativity, secure them, and combine this with the sequential executions. These algorithmic strategies are reinforced with programmability suggestions to assure deterministic executions. Finally, we demonstrate the applicability and the effectiveness of the proposed solutions as well as their performance on both CPUs and GPUs using the matrices from the SuiteSparse Matrix Collection, which show signs of the non-identical convergences and results.

This is a joint work with Daichi Mukunoki (R-CCS), Takeshi Ogita (TWCU), Katsuhisa Ozaki (Shibaura), and Stef Graillat (Sorbonne Université)

NAME **Martin KÜHN**

TITLE *Implicitly extrapolated geometric multigrid for the gyrokinetic Poisson equation from fusion plasma applications*

ABSTRACT In the context of Tokamak fusion plasma, the gyrokinetic Poisson equation has to be solved on disk-like domains which correspond to the poloidal cross-section of the Tokamak geometry;~see, e.g., [1,2]. In its simpliest form, this cross section takes a circular form but deformed geometries were found to be advantageous and are more realistic. We propose a taylored solver for the gyrokinetic Poisson equation on disk-like geometries described by curvilinear coordinates. Our solver also copes with an anisotropic, locally refined mesh to model the edge part of the Tokamaka and a rapidly dropping density profile as given by, e.g., [3,4].
Multigrid methods can achieve optimal complexity for many problems and are among the most efficient solvers for elliptic model problems. Multigrid methods for geometries described by curvilinear coordinates are however less common. We present a taylored geometric multigrid algorithm using optimized line smoothers to enable parallel scalability. We propose an implicit extrapolation scheme for our algorithm to increase the order of convergence by using nonstandard numerical integration rules for $\mathcal{P}_1$ finite elements. We also use problem-specific finite differences, which numerically show the same order of convergence as their finite elements' counterpart, enabling a matrix-free implementation with a low memory footprint.
[1] N. Bouzat, C. Bressan, V. Grandgirard, G. Latu, and M. Mehrenberger, Targeting realistic geometry
in tokamak code gysela, ESAIM: ProcS, 63 (2018), pp. 179–207.
[2] M. J. Kuhn, C. Kruse, and U. R ¨ ude, ¨ Implicitly extrapolated geometric multigrid on disk-like domains for the gyrokinetic Poisson equation from fusion plasma applications.
[3] E. Sonnendrucker, ¨ Private communication, 2019.
[4] E. Zoni, Theoretical and numerical studies of gyrokinetic models for shaped tokamak plasmas, PhD thesis, Technische Universitat München, München, 2019

NAME **Sherry LI**

TITLE *Leveraging One-Sided Communication for Sparse Triangular Solvers*

ABSTRACT We implement and evaluate a one-sided communication-based distributed-memory sparse triangular solve (SpTRSV). SpTRSV is used in conjunction with Sparse LU to affect preconditioning in linear solvers while one-sided communication paradigms enjoy higher effective network bandwidth and lower synchronization costs compared to their two-sided counterparts. We use a passive target mode in one-sided communication to implement a synchronization-free task queue to manage the messaging between producer-consumer pairs. Whereas some numerical methods lend themselves to simple performance analysis, the DAG-based computational graph of SpTRSV demands we construct a critical path performance model in order to assess our observed performance relative to machine capabilities. In alignment with our model, our foMPI-based one-sided implementation of SpTRSV reduces communication time by 1.5x to 2.5x and improves SpTRSV solver performance by up to 2.4x compared to the SuperLU_DIST's two-sided MPI implementation running on 64 to 4,096 processes on Cray supercomputers.

This is joint work with Nan Ding, Yang Liu, Sam Williams of LBNL.

NAME **Jennifer LOE**

TITLE *Multiprecision GMRES in Belos and Kokkos*

ABSTRACT We explore how changing precisions from double to float in some computations can help to speed up GMRES solves in the Trilinos package Belos. A Kokkos-based linear algebra adapter to Belos allows us to run on GPUs. We study converence for a variation of GMRES with iterative refinement and compare speedups for the specific linear algebra kernels contained therein. Furthermore, we study polynomial and block Jacobi preconditioning in float and double precision to see what performance can be attained on GPUs.

NAME **Robert LUCAS**

TITLE *Conveyors, an Abstraction for Message Aggregation*

ABSTRACT Beyond the end of Dennard scaling, and nearing the end of Moore’s Law, increased computing power is achieved by employing ever more processors, and using message passing to communicate amongst them. A familiar problem for those working with sparse matrices is the desire to send many small objects amongst the processors, such as tuples containing row and column indices, and a coefficient when redistributing a sparse matrix. The overhead of sending such small messages is prohibitive, forcing developers to aggregate them into larger buffers, which might then be exchanged using the MPI_ALLTOALLV function. Conveyors are a new mechanism for abstracting message aggregation. Data being sent is pushed to the conveyor library, while data to be received is pulled. All of the buffer management for aggregating the small messages, as well as the details of their communication, are managed by the library.

NAME **Mohsen MAHMOUDI AZNAVEH**

TITLE *Parallel multifrontal sparse LU factorization based on UMFPACK*

ABSTRACT We are introducing a new parallel sparse factorization algorithm that is based on a prior successful right-looking LU package, UMFPACK. UMFPACK is a part of the SuiteSparse package and appears as a built-in routine in MATLAB. We are designing a new algorithm from scratch that is amenable to parallelism which is a multifrontal algorithm like UMFPACK. The new algorithm is right-looking and uses the sparsest pivot like UMFPACK, however it uses different data structures which are coarse-grained and more suitable to call parallel BLAS kernels on bigger matrices. To avoid data dependency we have used different techniques; For example, we keep the list of columns of each front sorted and use a heap data structure to assemble pivotal columns. Also, we have a hash data structure of columns to be able to find columns of current front in nearly constant time. While there is no data dependency between the rows, our current design looks like UMFPACK for the rows. Without data dependencies among fronts we can have several fronts running together in parallel. We also will take advantage of parallelism within a front.
Mohsen Aznaveh and Tim Davis.

NAME **Pierre MATALON**

TITLE *Toward fast and accurate solutions of elliptic equations with HHO discretizations and multigrid methods*

ABSTRACT Hybrid High-Order (HHO) discretizations have gained growing interest in recent years for the solution of Partial Differential Equations. It applies to polytopal meshes, allows high orders of approximations and achieves optimal orders of convergence even on distorted meshes. The degrees of freedom (DoF) are located in cells and on faces and globally represent broken polynomials. The cell-DoFs being only locally coupled, their local elimination from the linear system gives rise to a Schur complement of smaller size where only face-DoFs remain. This process is known as static condensation, and the resulting system as a statically condensed or trace system. Our contribution proposes an efficient multigrid method to solve such large sparse trace systems for any arbitrary degree of approximation. A geometric approach is adopted, imposing very few constraints on the grid hierarchy. Indeed, the method is successfully applied to unstructured polytopal meshes, and the hierarchy does not have to be nested. However, it is required to coarsen not only the elements, but also the faces. To meet this requirement, a suited coarsening strategy is also proposed. Numerical tests are presented in 2D and 3D on the diffusion equation.

NAME **Francesco MEZZADRI**

TITLE *Iterative solution of horizontal linear complementarity problems on parallel computers*

ABSTRACT The horizontal linear complementarity problem (HLCP) is a well-known generalization of the linear complementarity problem (LCP). The HLCP has also many applications in various fields of science and engineering, and classical solution strategies include interior point methods or reducing the HLCP to a standard LCP. Recently, however, more direct approaches have been proposed as well, such as projected and modulus-based matrix splitting methods. These strategies are characterized by a simple iteration, which can

easily exploit the sparsity of the matrices of the problem. This is crucial especially in the case of large HLCPs of sparse matrices, where also parallel computing becomes interesting.
Thus, in this talk, we discuss recent developments on the parallel solution of HLCPs by matrix splitting techniques, devoting particular attention to the recently proposed modulus-based multisplitting methods for HLCPs. In this regard, first we introduce the problems and the methods of our concern, contextualizing them in the literature. We then outline the convergence of the methods and, finally, discuss the results of some numerical experiments.

NAME **Anastasiia MINENKOVA**
TITLE *Stability of Certain Canonical Forms under Small Perturbation*
ABSTRACT In this talk we discuss the behavior of some canonical forms like the Jordan form, the Weierstrass form, the Schur form under small structure preserving perturbations.

NAME **Ani MIRACI**
TITLE *Contractive local adaptive smoothing based on Dörfler's marking in a-posteriori-steered p-robust multigrid solvers*
ABSTRACT In this work we study a local adaptive smoothing algorithm for a-posteriori-steered p-robust multigrid methods. The solver tackles a linear system which is generated by the discretization of a second order elliptic diffusion problem using conforming finite elements of polynomial order $p \geq 1$. After one V-cycle ("full-smoothing" substep) of the solver of [HAL Preprint 02494538, 2020], we dispose of a reliable, efficient, and localized estimation of the algebraic error. We use this existing result to develop our new adaptive algorithm: thanks to the information of the estimator and based on a bulk chasing criterion, cf. Dörfler [SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124], we mark patches of elements and levels with increased estimated error. Then, we proceed by a modified and cheaper V-cycle ("adaptive-smoothing" substep), which only applies smoothing in the marked regions. The proposed adaptive multigrid solver picks autonomously and adaptively the optimal step-size per level as in our previous work but also the type of smoothing per level (weighted restricted additive or additive Schwarz) and concentrates smoothing to marked regions with high error. We prove that each substep (full and adaptive) contracts the error p-robustly, which is confirmed by numerical experiments. Moreover, the proposed algorithm behaves numerically robustly with respect to the number of levels as well as to the diffusion coefficient jump.

NAME **Ron MORGAN**
TITLE *Stable Polynomial Preconditioning for Linear Equations and Eigenvalues*
ABSTRACT Polynomial preconditioning can significantly improve Krylov solution of large systems of linear equations and large eigenvalue problems. Many past polynomial preconditioners have suffered from either difficult implementation or instability. We discuss a polynomial preconditioner that is based on the GMRES polynomial and is fairly simple to implement. We also give stability control that involves adding roots to the polynomial. Examples demonstrate that for difficult problems, the new approach can give remarkable

improvement in efficiency. Parallel computations can particularly benefit from the reduction of communication-intensive operations.

NAME **Erick MORENO-CENTENO**

TITLE *Exactly Solving Linear Systems via the Sparse Exact (SPEX) Factorization Framework*

ABSTRACT Solving sparse linear systems is fundamental for a plethora of algorithms and applications. Specifically, matrix factorizations play a central role in solving linear programs and several other optimization problems. Exactly solving (sparse) linear programs in particular, and linear systems, in general, is necessary for some applications (e.g., theoretical results, feasibility problems, military applications, applications involving hefty costs or profits, applications where ill-conditioning is common, etc.). To address this, we are developing the Sparse Exact (SPEX) Factorization Framework (currently comprising LU and Cholesky factorizations): a high-performance, well-documented, and extremely-robust software package. A key property of our SPEX factorizations is that the maximum word-length is polynomially bounded without using rational numbers or costly gcd calculations. Moreover, our SPEX Cholesky factorization solves SPD linear systems in time proportional to arithmetic work (i.e., has no data structure overhead). We benchmark the performance of our SPEX factorizations and the accuracy of MATLAB sparse backslash. The SPEX Factorization Framework is written in ANSI C, comes with a MATLAB interface, and is dually distributed via GitHub and as a component of SuiteSparse. This research is collaborative with Christopher Lourenco, Jinhao Chen, and Timothy Davis.

NAME **Esmond NG**

TITLE *Does the TSP heuristic for minimizing block counts in sparse Cholesky factorization have to be expensive?*

ABSTRACT In 2017, Pichon, Faverge, Ramet, and Roman introduced a method for reordering columns within supernodes of a sparse Cholesky factor to minimize the number of dense blocks. The method is based on reformulating the central optimization problem as a traveling salesman problem (TSP). The primary issue with their approach is its cost in time, and the primary bottleneck is computing the TSP distances. In this talk, we discuss techniques that reduce the time complexity of their method.

NAME **Grégoire PICHON**

TITLE *Trading Performance for Memory in Sparse Direct Solvers using Low-rank Compression*

ABSTRACT Sparse direct solvers using Block Low-Rank compression have been proven efficient to solve problems arising in many real-life applications. Improving those solvers is crucial for being able to 1) solve larger problems and 2) speed up computations. A main characteristic of a

sparse direct solver using low-rank compression is when compression is performed. There are two distinct approaches: (1) all blocks are compressed before starting the factorization, which reduces the memory as much as possible, or (2) each block is compressed as late as possible, which usually leads to better speedup. In this talk, we will present a composite approach, to speedup computations while staying under a given

memory limit. This allows to solve large problems that cannot be solved with Approach 2 while reducing the execution time compared to Approach 1. We propose a memory-aware strategy where each block can be compressed either at the beginning or as late as possible. We first consider the problem of choosing when to compress each block, under the assumption that all information on blocks is perfectly known, i.e., memory requirement and execution time of a block when compressed or not. We show that this problem is a variant of the NP-complete Knapsack problem, and adapt an existing 2-approximation algorithm for our problem. Unfortunately, the required information on blocks depends on numerical properties and in practice cannot be known in advance. We thus introduce models to estimate those values. Experiments on the PaStiX solver demonstrate that our new approach can achieve an excellent trade-off between memory consumption and computational cost. For instance on matrix Geo1438, Approach 2 uses three times as much memory as Approach 1 while being three times faster. Our new approach leads to an execution time only 30% larger than Approach 2 when given a memory 30% larger than the one needed by Approach 1.

NAME **Jason RIEDY**

TITLE *Graph Analysis and Novel Architectures*

ABSTRACT The memory and threading systems in current computer architectures do not support the kinds of sparsity needed for highly efficient graph analysis. Cache lines and bandwidth are partially utilized. Processors sit idle. New techniques targeting the memory access and parallelism requirements for sparse graph and matrix analysis are proposed and being deployed in novel computer architectures. We discuss proposed architectures and their wider implications. We present details on one physically realized architecture based on migrating threads with fine-grained memory access along with its GraphBLAS implementation.

NAME **Jemima TABEART**

TITLE *Preconditioners for saddle point weak-constraint 4D-Var with correlated observation errors*

ABSTRACT Data assimilation algorithms for numerical weather prediction are increasingly high dimensional, and complicated with the widespread use of correlated observation error covariance matrices. Fast convergence is essential, making the saddle point formulation of weak-constraint 4D-Var desirable due to its potential for parallelization. We consider new preconditioners which better approximate the model and observation error terms in the inexact constraint preconditioner. Previous work has found that correlated observation errors, which are often neglected by standard preconditioners, can lead to ill-conditioned data assimilation problems. We show for the heat equation that accounting for observation error correlations in the preconditioners is necessary for fast convergence. Our new approach results in faster convergence in serial than standard approaches, and can be easily adapted to exploit parallel computer architectures.

NAME **Roel TIELEN**

TITLE *Block ILUT smoothers for p-multigrid methods in Isogeometric Analysis*

ABSTRACT Isogeometric Analysis [1] has become increasingly popular as an alternative to the Finite Element Method. Solving the resulting linear systems when adopting higher order B-spline basis functions remains a challenging task, as most (standard) iterative methods have a deteriorating performance for higher values of the approximation order p.
Recently, we successfully applied p-multigrid methods to discretizations arising in Isogeometric Analysis [2]. In contrast to h-multigrid methods, where each level of the multigrid hierarchy corresponds to a different mesh width h, the p-multigrid hierarchy is constructed based on different approximation orders. The residual equation is then solved at level p = 1, where B-spline basis functions coincide with standard Lagrangian P 1 basis functions. This enables the use of efficient solution techniques developed for low-order standard FEM. Numerical results, for both single patch and multipatch geometries show, that the number of iterations needed for convergence is independent of both h and p when the p-multigrid method is enhanced with a smoother based on an Incomplete LU factorization with dual threshold (ILUT). However, a slight dependence on the number of patches has been observed for multipatch geometries.
Since the resulting system matrix has a block structure in case of a multipatch geometry, we consider the use of block ILUT as a smoother. Results indicate that the use of block ILUT can be an efficient alternative to ILUT on multipatch geometries within a heterogeneous HPC framework. Preliminary results for p-multigrid methods adopting a block ILUT smoother will be presented in this talk.

References
[1] T.J.R. Hughes, J.A. Cottrell and Y. Bazilevs, Isogeometric Analysis: CAD, Finite Elements, NURBS, Exact Geometry and Mesh Refinement, Computer Methods in Applied Mechanics and Engineering, 194, 4135 - 4195, 2005
[2] R.Tielen, M. Möller, D. Göddeke and C. Vuik, p-multigrid methods and their comparison to h-multigrid methods within Isogeometric Analysis, Computer Methods in Applied Mechanics and Engineering, 372, 2020

NAME **Wim VANROOSE**

TITLE *Krylov-Simplex method to solve inverse problems in l1-norm and max-norm.*

ABSTRACT A Krylov method solves a problem in a subspace that is expanded, each iteration, by one additional sparse matrix-vector product. When an optimisation problem is projected on such a Krylov subspace this leads to a small and easy to solve problem. For example GMRES leads to a small least-squares problem. Let us solve min ‖Ax−b‖_∞, for a non-square matrix with a Krylov subspace. The projected problem can be reformulated as a LP problem, for a given a basis V_k for K_k (A^T A, A^Tr_0) generated, for example, by Golub-Kahan. We then design a simplex method to solve the projected problem. We are able to update the basic set as the Krylov method expands and update the LU factorisation from the previous simplex step, exporting the properties of the Krylov subspace.
In a similar way, it is possible to solve l1-optimization problems by a combination of Krylov and simplex iterations.
The talk give an overview of the initial steps of this research that has many applications in inverse problems.

NAME **Bastien VIEUBLÉ**

TITLE *Multiple precisions iterative refinement for the solution of large sparse linear systems*

ABSTRACT The increasing availability of half precision (fp16, bfloat16) in hardware tends to lead modern high performance computation to use multiple precision standards. In that context iterative refinement (IR), originally focused on improving the accuracy of linear system solution, has been embracing low precision computation. In particular Carson and Higham in their recent papers present a three precisions version of the IR which has been proven highly successful at solving dense linear systems at high performance and with high accuracy. They also developed a GMRES-based iterative refinement (GMRES-IR) which can handle ill-conditioned matrices even when the LU factorization is carried out in low precision. In this talk, we discuss the usability and potential limits of these techniques for the solution of large scale sparse linear systems arising in real life applications. Based on these observations, we propose a generalized GMRES-IR using up to five precisions, we present the associated error analysis and discuss the set of relevant precision combinations which potentially improve either the accuracy and robustness or the performance (memory consumption and execution time) of the solver. Our experimental results, based on the MUMPS multifrontal solver, demonstrate the effectiveness of the proposed method, which can achieve significant gains in both memory consumption and performance on parallel machines and/or provide accurate solutions for numerically difficult problems.