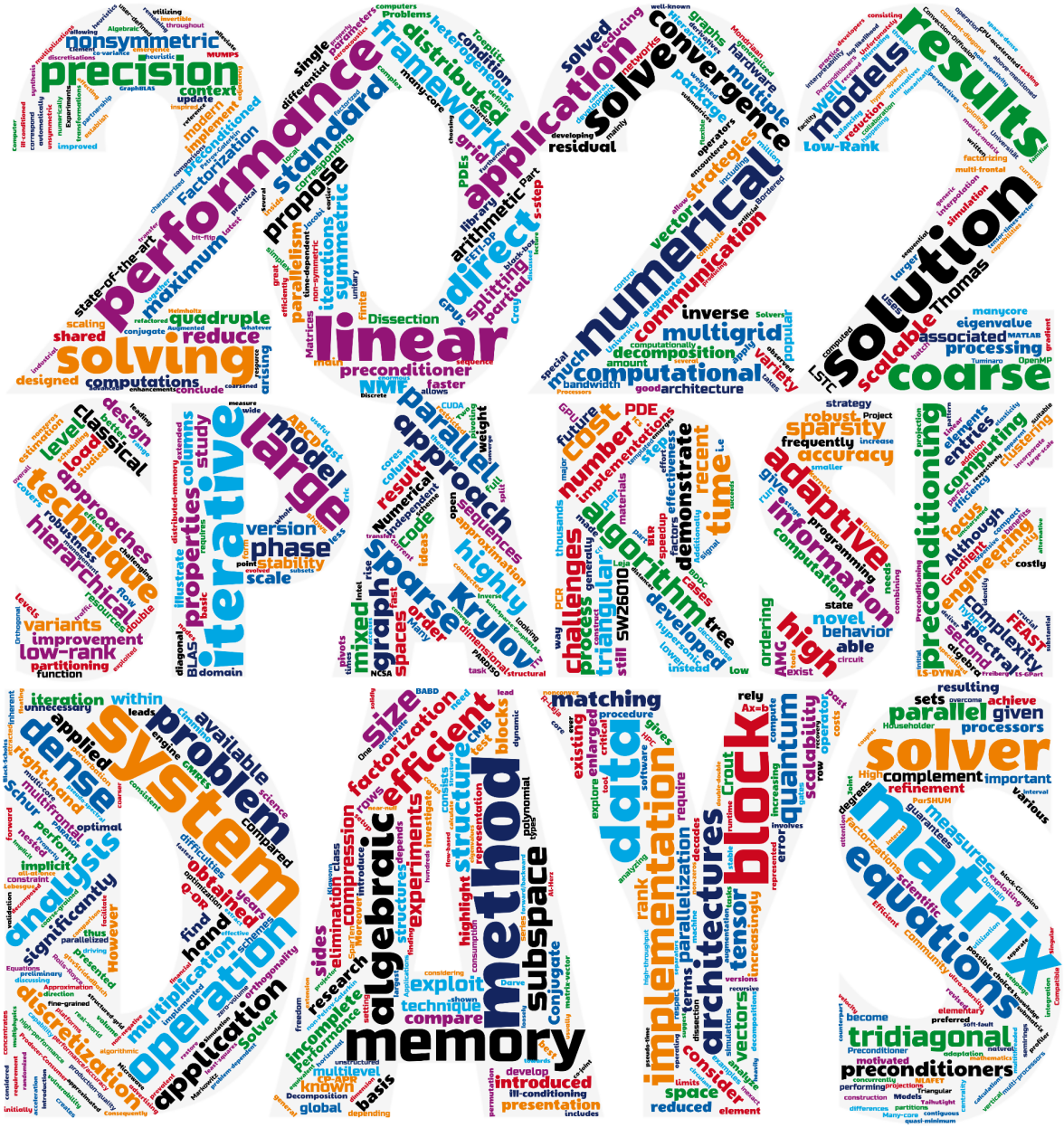

Sparse days meeting 2022 in Saint-Girons

CERFACS · ANITI · CIMI TOULOUSE · IRIT · N7 DÉVELOPPEMENT

Saint Girons · Ariège 09 · Atos · Hewlett Packard Enterprise · NEC

Abstracts

SESSION 1 — Sparse direct methods, combinatorics, graphs I

Towards Two to Five Truths Revealed in Non-Negative Matrix Factorizations

John Conroy

In this talk we explore the role of matrix scaling of a matrix of counts when building a topic model using a non-negative matrix factorization. We present a scaling inspired by the normalized Laplacian (NL) for graphs can greatly improve the quality of a non-negative matrix factorization. The results parallel those in spectral graph clustering work of [2], where the authors proved adjacency spectral embedding (ASE) spectral clustering was more likely to discover core-periphery partitions and Laplacian Spectral Embedding (LSE) was more likely to discover affinity partitions. In the text applications non-negative factorizations of matrices (NMF) are typically used on a matrix of co-occurrence “contexts” and “terms” counts. 1 Terms, in our experiments were tokens as defined by CountVectorizer, which breaks on white space and removes punctuation. Depending on the application a context could be a set of two or more consecutive terms, a sentence, or one or more documents. We studied five matrix scalings (1) the original counts (None), (2) column scaling (CS), (3) row scaling (RS), (4) pointwise mutual information (PWMI), and (5) normalized Laplacian (NL). As the matrices are context-term matrices RS scaling turn each row into the maximum likelihood multinomial for context and CS estimates the term distributions across the context. PWMI scales the counts by the row and column marginals, while NL scales them by the square root's of these marginals. To the extent that the spectral result of [2] carries over to NMF applied to text we would expect NL to be best at recovering latent topics. We test these 5 scalings on three datasets, 20 newsgroups, NMF both ways using counts and LSE normalized counts for the newsgroup data (20 groups), Clemson Russian troll data (8 types of trolls). 2 and a collection of 500 newswire documents, consisting of 50 topics from the Document Understanding Conference (DUC). Using the adjusted Rand index (ARI) measure cluster similarity we see an increase of 50% for the tweet data and over 200% for the newsgroup dataset. For the DUC dataset, NL edges out RS and gives over 40% improvement over None. The second part of the talk will give some analysis of why these matrix scalings behave as they do.

[1] John M. Conroy, Sashka T. Davis, Section mixture models for scientific document summarization International Journal on Digital Libraries 19 (2-3), 305-322, 2018.

[2] 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.

Bale: A discussion of aggregating communication for parallel programming

Jason Devinney, John Gilbert

In 2018, we released a suite of C software called bale on GitHub. This software suite is not a tool or a benchmark. So what is bale? Bale is an effort to spark discussion about parallel programming productivity, specifically concerning irregular algorithms and communication aggregation. Bale

includes our attempts at aggregation libraries and several mini-kernels that exercise those libraries. We think communication aggregation, once viewed as a necessary evil, is actually a key element of current and future parallel programming. However, it is often very difficult to write code that aggregates communications, even with our libraries. We have been exploring new ways to incorporate aggregation into programming languages and we would like to include more people in the discussion.

SuiteSparse:GraphBLAS: parallel graph algorithms via sparse matrix computations over semirings

Tim Davis

SuiteSparse:GraphBLAS is a full parallel 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. A description of the parallel implementation of SuiteSparse:GraphBLAS is given, including its novel parallel algorithms for sparse matrix multiply, addition, element-wise multiply, submatrix extraction and assignment, and the GraphBLAS mask/accumulator operation. Its performance is illustrated by solving the graph problems in the GAP Benchmark and by comparing it with other sparse matrix libraries.

Scalable Partitioning of Directed Graphs Using Graphlets

Luce Le Gorrec, Philip Knight

Community detection aims to partition a network into similar groups of nodes. Mainstream approaches to perform this focus on direct interactions (i.e. edges), but this is limiting for many pattern-based community structures that can be met in directed networks, such as citation-based or flow-based communities [1]. Recently, analyses based on implicit interactions, or higher-order structures, have been investigated [2,3,4]. In particular, those based on small subgraphs, or graphlets, have been identified as a promising tool.

Here, we propose an algorithm to partition directed networks based on graphlets. Namely, our method uses graphlets to produce an undirected representation of the network called its Motif Adjacency Matrix (MAM) [2], and partitions this representation using the Louvain algorithm [5] (but it is designed so that other partitioning algorithms could be used instead). It finally uses a homebrewed adaptation of the Louvain algorithm to assign nodes disconnected from the MAM to some partition. Our algorithm implementation is **versatile**, in the sense that it addresses a large set of graphlets. While other existing methods produce MAM of essentially 3-node graphlets, our solution can also address quadrangles via an adaptation of the algorithm from [6]. Our implemented solution is also **numerically efficient**, as it addresses networks up to a few millions of nodes. On these both points, to the best of our knowledge, our implementation is the state-of-the-art solution.

The software is publicly available at https://github.com/luleg/PartitionMAM.

[1] Malliaros, F. D., & Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. Physics reports, 533(4), 95-142.

[2] Benson, A. R., Gleich, D. F., & Leskovec, J. (2016). Higher-order organization of complex networks. Science, 353(6295), 163-166.

[3] Underwood, W. G., Elliott, A., & Cucuringu, M. (2020). Motif-based spectral clustering of ewighted directed networks. Applied Network Science, 5(1), 1-41.

[4] Tsourakakis, C. E., Pachocki, J., & Mitzenmacher, M. (2017, April). Scalable motif aware graph clustering. In Proceedings of the 26th International Conference on World Wide Web (pp. 1451-1460)

[5] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.

[6] Chiba, N., & Nishizeki, T. (1985). Arboricity and subgraph listing algorithms. SIAM Journal on computing, 14(1), 210-223.

SESSION 2 — High performance computing I, quantum computing

Jack Dongarra	A Look at Mixed Precision Solver
Jean-Baptiste Harry	NEC SX-Aurora TSUBASA vector architecture with high memory bandwidth for the linear algebra solvers
Bob Lucas	Beyond GPUs
Marc Baboulin	Optimizing quantum algorithms using matrix factorization

A Look at Mixed Precision Solver

Jack Dongarra

There is a rapidly expanding landscape of mixed- and multi-precision methods. The ongoing cross-pollination between high-performance computing (HPC) and machine learning (ML) is leading to intelligent computational steering of large-scale simulations. Sharing of the hardware platforms and exploiting their wide range of computational modes has led to proliferation of multiple representations of floating-point data—and increased interest in new methods that exploit them. Against the backdrop of high-performance libraries produced by internet-scale companies, hardware vendors, national laboratories, and academic institutions, we will show the recent algorithmic progress in exploiting multiple precisions for increased efficiency in performance, communication, and/or storage.

NEC SX-Aurora TSUBASA vector architecture with high memory bandwidth for the linear algebra solvers

Jean-Baptiste Harry

The Japanese Company NEC Corporation is an historic actor of the HPC both in the classical x86 servers and the vector architectures. The usage of the vector architectures was in expansion until the 2000's in a mainframe form factor. The high point is the Earth Simulator supercomputer which was the world most powerful from 2001 to 2004.

The 2000's was the x86 Linux expansion years but since 2010's the accelerator share is rising, and the supercomputer are becoming more powerful and heterogenous. That's why NEC decided to provide to the market a more affordable vector technologies in an PCIe accelerator form factor in 2018.

Historically in meteorology and seismic, the accelerator form factor allows to generalize the use in new domains like fluid mechanic, structural mechanic, machine learning, deep learning, molecular dynamic and others.

The NEC SX-Aurora TSUBASA PCIe card is a vector system who is integrated to a x86 Linux server. The PCIe ports allow to connect up to 8 cards on each server. These cards can support Linux OS, it allows to use the card as an accelerator, a coprocessor, or a principal processor. This vector processor is built with 6 HBM2 modules for a fast memory access and a better efficiency on problems where memory bandwidth is the limiting factor.

This vector technology supports open standard languages (Fortran, C, C++, Python ...), open standard libraries (MPI, openMP, bas, lapack ...) and AI framework (Pytorch, TensorFlow, Spark ...). The objective is to keep a unique source code and runnable on all systems. During the presentation, we will explain NEC vector technologies and illustrate the application through: - synthetic benchmarks (Stream, HPL, HPCG) - collaborations and results in the domain of sparse and direct solvers - illustration on some real applications.

Beyond GPUs
Bob Lucas

The end of Dennard scaling and the growth in demand for machine learning have led to a diverse range of new computing devices that can be considered as accelerators for sparse matrix computations, beyond the Graphics Processing Units (GPUs) already familiar to this community. There are compute intensive devices such as the 850,000 core Cerebras CS-2, the World's largest chip, and racks full of Google Tensor Processing Units (TPUs). There are also novel data-flow based systems such as Samba Nova and Next Silicon. This talk will discuss two efforts to examine additional possible accelerators. The Xilinx Alveo compute card couples an FPGA with High Bandwidth Memory (HBM), offering an alternative for memory bandwidth bound computations such as a preconditioned conjugate gradients linear solver, the default solver for thermal calculations in LS-DYNA.. The D-Wave quantum annealer finds the ground state of a user supplied Ising model, an NP-complete problem, that ought to be able to perform the graph partitioning for a nested dissection reordering. We will present early results with both the Xilinx Alveo and the D-Wave Advantage.

Optimizing quantum algorithms using matrix factorization
Marc Baboulin

Quantum Computing aims at addressing scientific computations that are currently intractable by conventional supercomputers and to achieve the so-called “quantum supremacy” but it is also a promising technology for speeding up existing simulations. After a short introduction to Quantum Computing using a linear algebra approach, we illustrate how we can efficiently decompose a quantum operator using Householder transformations. This method is particularly useful for compiling a quantum program or encoding classical data into a quantum memory.

SESSION 3 — Iterative and hybrid solvers

Pierre Matalon	Algebraic multigrid for condensed systems arising from hybrid discretizations
Yongseok Jang	Randomized GMRES with Singular Vectors Based Deflated Restarting
Christie Louis Alappat	RACE: Speeding Up Iterative Solvers Using Level-Based Blocking Techniques
Alexis Montoison	Krylov.jl: A Julia basket of hand-picked Krylov solvers

Algebraic multigrid for condensed systems arising from hybrid discretizations
Pierre Matalon, Daniele Di Pietro, Frank Hülsemann, Paul Mycek, Ulrich Ruede

We address the numerical solution of linear systems arising from the hybrid discretizations of second-order elliptic partial differential equations. Such discretizations hinge on a hybrid set of degrees of freedom (DoFs), respectively defined in cells and faces, which naturally gives rise to a global hybrid system of linear equations. Assuming that the cell unknowns are only locally coupled, they can be efficiently eliminated from the system, leaving only face unknowns in the resulting Schur complement, also called statically condensed matrix. We propose an algebraic multigrid (AMG) preconditioner specifically targeting condensed systems corresponding to lowest order discretizations (piecewise constant). Like traditional AMG methods, we retrieve geometric

information on the coupling of the DoFs from algebraic data. However, as the condensed matrix only gives information on the faces, we use the uncondensed version to reconstruct the connectivity graph between elements and faces. An aggregation-based coarsening strategy mimicking a geometric coarsening or semi-coarsening can then be set up to build coarse levels. Numerical experiments are performed on diffusion problems discretized by the Hybrid High-Order (HHO) method at the lowest order. Our approach uses a K-cycle to precondition an outer flexible Krylov method. The results demonstrate similar performances, in most cases, compared to a standard AMG method, and a notable improvement on anisotropic problems with Cartesian meshes.

Randomized GMRES with Singular Vectors Based Deflated Restarting
Yongseok Jang, Cédric Content, Emeric Martin, Laura Grigori

For high dimensional spaces, a randomized Gram-Schmidt algorithm is beneficial in computational costs as well as numerical stability. We apply this dimension reduction technique by random sketching to Krylov subspace methods, e.g. to GMRES. We propose a flexible variant of GMRES with the randomized Gram-Schmidt based Arnoldi iteration to produce a set of basis vectors of the Krylov subspace. Even though the Krylov basis is no longer l2 orthonormal, its random projection onto the low dimensional space shows l2 orthogonality. As a result, it is observed the numerical stability which turns out to be independent of the dimension of the problem even in extreme scale problems. On the other hand, as the Harmonic Ritz values are commonly used in GMRES with deflated restarting to improve convergence, we consider another deflation strategy, for instance disregarding the singular vectors associated with the smallest singular values. We thus introduce a new algorithm of the randomized flexible GMRES with SVD based deflated restarting. At the end, we carry out some numerical experiments in the context of compressible turbulent flow simulations. Our proposed approach exhibits a quite competitive numerical behavior to existing methods while reducing computational costs.

RACE: Speeding Up Iterative Solvers Using Level-Based Blocking Techniques
Christie Louis Alappat, Georg Hager, Gerhard Wellein

Sparse linear iterative solvers are indispensable for many large-scale simulations. In this talk, we present methods to accelerate some of the existing solvers by using the concept of levels as developed in the context of our Recursive Algebraic Coloring Engine (RACE) library. Levels are constructed using breadth-first search (BFS) on the graph that represents the underlying sparse matrix. In kernels like sparse-matrix-power vector multiplication and Chebyshev polynomial iterations, which perform repetitive back-to back sparse-matrix-vector multiplication (SpMV)-type iterations without global synchronizations, these levels are then used to implement cache blocking of the matrix elements for high spatial and temporal reuse. The method is highly effective and achieves performance levels of 50-100 GF/s on a single modern Intel or AMD multicore chip, providing speedups of typically 2x-4x compared to a highly optimized standard SpMV implementation.

After introducing the optimization strategy, we demonstrate the benefits of such optimized kernels in iterative solvers. To this end, we discuss the coupling of the RACE library with the Trilinos framework and address the application to communication-avoiding s-step GMRES solver and polynomial preconditioners.

Krylov.jl: A Julia basket of hand-picked Krylov solvers
Alexis Montoison, Dominique Orban

Krylov.jl provides Julia implementations of some of the most useful Krylov solvers for linear systems, least-squares problems, least-norm problems, adjoint systems, and saddle-point systems.

Krylov solvers are appropriate when a factorization is not practical. For example, if

- the matrix is not available explicitly,
- the matrix or its factors would require excessive memory if stored.

The solvers in Krylov.jl have been optimized for time and memory, have an in-place version, are compatible with GPU, and work in any precision (real and complex numbers).

SESSION 4 — Artificial intelligence, machine learning I

Sherry Li	Autotuning sparse linear solvers and theirs applications with Gaussian process regression
Michela Taufer	A Suite Of AI-based Tools For IO-aware HPC Resource Management
Michael Kirby	An Autoencoder Neural Network Architecture for Supervised Sparse Nonlinear Principal Component Analysis

Autotuning sparse linear solvers and theirs applications with Gaussian process regression
Sherry Li

Significant effort has been invested to develop highly scalable numerical libraries and high-fidelity modeling and simulation for the upcoming exascale computers. These codes typically involve many parameters that influence performance on the underlying parallel machine. They are also expensive to run and thus have limited “function evaluation” values, which post significant challenges to efficient performance tuning on diverse architectures.

Bayesian optimization with Gaussian process regression is an attractive machine learning framework to build surrogate models with limited function evaluation points. In order to fully utilize all the available data, we leverage multitask learning and multi-armed bandit strategies to build a more advanced Bayesian optimization framework. We will show several advanced features of GPTune, e.g., incorporation of coarse performance models to improve the surrogate model, multi-objective tuning such as tuning a hybrid of time, memory and accuracy, and use of historical database for transfer learning. We will demonstrate the efficiency and effectiveness of GPTune when it is applied to sparse linear algebra libraries, such as SuperLU and Hypre, as well as production-level fusion simulation codes, such as NIMROD.

AI4IO: A Suite Of AI-based Tools For IO-aware HPC Resource Management
Michela Taufer

High performance computing (HPC) is undergoing many changes at the system level. While scientific applications can reach petaflops or more in computing performance, potentially resulting in larger data generation rates and more frequent checkpointing, the data movement to the parallel file system remains costly due to constraints imposed by HPC centers on the IO bandwidth. In other words, the bandwidth to file systems is outpaced by the rate of data generation; the associated IO contention increases job runtime and delays execution. This situation is aggravated by the fact that when users submit their jobs to a HPC system, they rely on resource managers and job schedulers to monitor and manage the computing resources (i.e., nodes). Both resource managers and job schedulers remain blind to the impact of IO contention on the overall simulation performance.

In this talk we discuss how Artificial Intelligence (AI) can augment HPC systems to prevent and mitigate IO contention while dealing with IO bandwidth constraints. Our solution, called Analytics for IO (AI4IO), consists of a suite of AI-based tools that enable IO-awareness on HPC systems. Specifically, we present two AI4IO tools: PRIONN and CanarIO. PRIONN automates predictions about user-submitted job resource usage, including per-job IO bandwidth; CanarIO detects, in real-time, the presence of IO contention on HPC systems and predicts which jobs are affected by that contention (e.g., because of their frequent checkpointing). By working in concert, PRIONN and

CanarIO predict the a priori knowledge necessary to prevent and mitigate IO contention with IO-aware scheduling. We integrate AI4IO in the Flux scheduler and show how A4IO produce improvements in simulation performance: we observe up to 6.2% improvement in makespan of HPC job workloads, which amounts to more than 18,000 node-hours saved per week on a production-size cluster. Our work is the first step to implementing IO-aware scheduling on production HPC systems.

An Autoencoder Neural Network Architecture for Supervised Sparse Nonlinear Principal Component Analysis

Michael Kirby, Tomojit Ghosh

Neural Networks in the form of autoencoders have been widely used for nonlinear principal component analysis. This framework for model and data reduction is still widely used today. We have proposed a modification to the idea of nonlinear autoencoders when data class labels are available, i.e., a form of supervised nonlinear PCA. This framework is useful for data visualization and classification. In this talk we will discuss a machine learning problem for the extraction of data features that are most useful for discriminating amongst a set of classes. Mathematically, a sparse optimization problem formulation is used for the determination of a minimal set of features that discriminate two or more classes indexed by the set. These centroid encoder neural networks map data through a space of reduced dimension to the data class centroid in the original space. The loss function of sparse centroid-encoder is defined as the distortion error with a hyper-parameter which is used to adjust the level of sparsity promoted by the one-norm penalty on thr sparsity paramters. There is an encoder function that is responsible for dimension reduction and a decoder function that reconstructs the data. This is a sparse implementation of the centroid-encoder for nonlinear data reduction and visualization that we refer to as Sparse Centroid-Encoder (SCE). We will demonstrate that this approach can produce state-of-the-art benchmarking results for feature selection across a range of data sets including, single-cell biological data, high dimensional infectious disease data, hyperspectral data, image data, and speech data. One key attribute of SCE is that it can extract informative features from a multi-modal data set, i.e., data sets whose classes appear to have multiple clusters. More broadly, it appears that the development of optimization algorithms that promote sparsity, for both convex and non-convex problems, are very attractive tools for learning from data.

SESSION 5 — Sparse direct methods, combinatorics, graphs II

Philip Knight	Scaling to semi-doubly stochastic form
Esmond Ng	Some observations regarding high-performance serial implementations of sparse symmetric factorization
Elisa Riccietti	Sparse matrix factorization from an optimization point of view

Scaling to semi-doubly stochastic form

Philip Knight, Luce le Gorrec, Daniel Ruiz, Sandrine Mouysset

The Sinkhorn-Knopp algorithm (SKA) is a very well known way of scaling a square nonnegative matrix into a doubly stochastic matrix. SKA, and (faster) variants, have proven to be useful in a wide variety of applications (machine learning, bioinformatics, matching, etc). While it is impossible for a non-square nonnegative matrix to have row and column sums both equal to one, SKA can be adapted to scale rectangular matrices to prescribed row and column sums, assuming such scalings exist. We present a new generalisation of doubly stochastic matrices with particular relevance to

rectangular matrices, which we christen “semi-doubly stochastic” (SDS). A nonnegative matrix A is SDS if A^TA and AA^T are doubly stochastic.

In contrast to doubly stochastic matrices, SDS matrices need not have support. We look at the implications in terms of the block structure of SDS matrices and show that so long as a rectangular matrix has the same nonzero pattern as an SDS matrix then it can be scaled to SDS form. While the block structure of semi-scalable matrices looks attractive, as this clearly has applications in co-clustering, there is no easy way to tell a priori whether a matrix has this property or not. In practice, if we attempt to use current scaling algorithms on such matrices without pre-existing knowledge of the underlying block structure, then they will fail to converge to anything meaningful. To remedy this, we present a new iterative scaling algorithm related to SK. We show the potential of a variant based on a Newton iteration. For an efficient implementation of the Newton step we need to solve a singular system implicitly to find a minimum norm least squares solution. We discuss techniques for circumventing potential pitfalls with this approach.

Some observations regarding high-performance serial implementations of sparse symmetric factorization

Esmond Ng, Barry Peyton

We consider the serial factorization of a sparse symmetric matrix into triangular factors. It is well known that sparse matrix factorizations incur fill, which means that there are typically more nonzero elements in the triangular factors than in the original matrix. For efficient implementations, the algorithms are designed so that only the nonzero elements are stored and manipulated. The performance of a sparse factorization algorithm is measured by the time and storage required by the computation, which, in turn, depend on the number of nonzero elements in the factors, their locations in the factors, the data structures used to store the nonzero elements, and the way these nonzero elements are manipulated. In this talk, we revisit these issues and discuss a number of implementations, including the left-looking, right-looking, and multifrontal approaches. We will conclude with some interesting observations.

Sparse matrix factorization from an optimization point of view

Quoc Tung Le, Elisa Riccietti, Rémi Gribonval

Approximating a dense matrix by a product of sparse factors is a fundamental problem for many signal processing and machine learning tasks. It can be formulated as a constrained optimization problem and decomposed into two subproblems: finding the positions of the non-zero coefficients in the sparse factors, and determining their values. While the first step is usually seen as the most challenging one due to its combinatorial nature, this talk focuses on the second step, referred to as sparse matrix approximation with fixed support (FSMF). We show the NP-hardness of FSMF, we present a non-trivial family of support constraints making the FSMF problem practically tractable with a fast gradient-free algorithm, and we investigate the landscape of the FSMF optimization formulation, proving the absence of spurious local valleys and spurious local minima.

SESSION 6 — High performance computing II

Somesh Singh	An Efficient Parallel Implementation of a Perfect Hashing Method for Hypergraphs
Joseph Touzet	A Large-Scale Distributed Simulation Framework for Irregular Quantum Dynamics
Ewa Deelman	Living in a Heterogenous World: How scientific workflows bridge diverse cyberinfrastructure and what we can do better?
Dzenek Dostal	Unpreconditioned hybrid TFETI methods for huge elliptic problems
Antoine Jego	Task-Based Parallel Programming for Scalable Algorithms

An Efficient Parallel Implementation of a Perfect Hashing Method for Hypergraphs
Somesh Singh, Bora Uçar

Querying the existence of an edge in a given graph or hypergraph is a building block in several algorithms. Hashing-based methods can be used for this purpose, where the given edges are stored in a hash table in a preprocessing step, and then the queries are answered using the lookup operations. While the general hashing methods have fast lookup times in the average case, the worst case run time is much higher. Perfect hashing methods take advantage of the fact that the items to be stored are all available and construct a collision free hash function for the given input, resulting in an optimal lookup time even in the worst case. We investigate an efficient shared-memory parallel implementation of a recently proposed perfect hashing method for hypergraphs. We experimentally compare the resulting parallel algorithms with the state-of-the-art and demonstrate better run time and scalability on a set of hypergraphs corresponding to real-life sparse tensors.

A Large-Scale Distributed Simulation Framework for Irregular Quantum Dynamics
Joseph Touzet, Kaya Oguz, Pablo Arrighi, Amélia Durbec

In traditional quantum computing, e.g. in the quantum circuit model, the size of the data structure describing basis elements is well known, because the dimensionality is fixed. General quantum systems, however, exhibit basis elements of variable sizes, and state spaces having dynamically unbounded, possibly infinite dimensionality, e.g. for quantum Turing machines or quantum field theories. In this paper, we propose a computational framework for a scalable simulation of such general irregular quantum systems in distributed memory parallel environments. We lay out the computational challenges arising from the nature of such simulations, then propose effective parallelization, load balancing as well as memory management strategies to accelerate them. We report scalability results for up to 1584 MPI processes on a parallel cluster using our framework for the special case of a quantum causal graph dynamics simulation.

Living in a Heterogenous World: How scientific workflows bridge diverse cyberinfrastructure and what we can do better?
Ewa Deelman

Scientific workflows are now a common tool used by domain scientists in a number of disciplines. They are appealing because they enable users to think at high level of abstraction, composing complex applications from individual application components. Workflow management systems (WMSs), such as Pegasus (http://pegasus.isi.edu) automate the process of executing these workflows on modern cyberinfrastructure. They take these high-level, resource-independent descriptions and map them onto the available heterogeneous resources: campus clusters, high-performance computing resources, high-throughput resources, clouds, and the edge. WMSs can select the appropriate resources based on their architecture, availability of key software,

performance, reliability, availability of cycles, storage space, among others. Using algorithms like those used in compilers, they can determine what data to save during execution, and which are no longer needed. Similarly to compiler solutions, they can generate an executable workflow that is tailored to the target execution environment, taking into account reliability, scalability, and performance. WMS use workflow execution engines to run the executable workflows on the target resources, while the jobs within the workflow are managed by the host runtime system. This talk will describe the key concepts used in the Pegasus WMS and pose the question whether there is a need and a desire to build systems in which WMS and compilers, and workflow execution engines and schedulers/runtime systems operate in tandem to deliver robust solutions to the scientists. Supporting tighter integration of workflow management and schedulers would potentially improve the performance of applications and system utilization.

Unpreconditioned hybrid TFETI methods for huge elliptic problems
Zdenek Dostal

The scope of scalability of classical FETI and FETI-DP methods is limited by the dimension of the coarse problem, which is proportional to the number of subdomains. The coarse problem is typically solved by a direct solver - its cost is negligible for a few subdomains, however, it starts to dominate when the number of subdomains is large, currently tens of thousands of subdomains. To overcome the latter limitation, Klawonn and Rheinbach proposed to enforce some constraints on the primal level by interconnecting the groups of subdomains into clusters. The defect of each cluster is then the same as that of each of its subdomains. Klawonn and Rheinbach coined the method H-FETI (hybrid FETI, also H-FETI-DP). Here we consider unpreconditioned variant H-TFETI (hybrid total) method which enforces the Dirichlet conditions by Lagrange multipliers. We present the analysis of conditioning of the dual matrices of the scalar and elastic model problems in 2D and 3D and results of numerical experiments with the solution of linear systems with billions of unknowns. We show, e.g., that the condition number of the cluster defined on a fixed cube domain decomposed into $m \times m \times m$ subdomains interconnected by face averages and discretized by regular grid increases proportionally to m including realistic bounds on relevant constraints. The results of numerical experiments show the considerable scope of scalability of both H-TFETI and indicate that this method can be helpful for the solution of huge linear and contact problems. The latter is not surprising as the estimate indicates that the cost of iterations increases proportionally to at most square of m, but the cost of the coarse problem decreases with the sixth power of m. Moreover, the two-level structure of the coarse grid can be effectively exploited by the node-core structure of the modern supercomputers' hardware.

Task-Based Parallel Programming for Scalable Algorithms
Antoine Jego, Alfredo Buttari, Emmanuel Agullo, Abdou Guermouche, Julien Herrmann

Task-based programming models have succeeded in gaining the interest of the high-performance mathematical software community thanks to how they relieve part of the burden of developing and implementing distributed-memory parallel algorithms in an efficient and portable way. In increasingly larger, more heterogenous clusters of computers, these models appear as a way to maintain and enhance more complex algorithms. However, task-based programming models lack flexibility and features that are necessary to express in an elegant and compact way scalable algorithms. We show that the Sequential Task Flow paradigm can be extended to write in a compact yet efficient and portable way scalable linear algebra algorithms such as the General Matrix Multiplication or the Cholesky factorization of dense matrices and their 2.5 or 3D variants. This extension required few modifications to the StarPU runtime system. The achieved implementations are shown to be competitive up to 32,768 cores with state-of-the-art libraries and may outperform them on some specific problem configurations.

SESSION 7 — Low rank approximation, variable precision, randomization I

Laura Grigori	Randomization for solving linear systems and eigenvalue problems
Theo Mary	Adaptive Precision Solvers for Sparse and Data Sparse Systems
Edmond Chow	H2Pack: Software for H2 Hierarchical Matrices Using the Proxy Point Method
George Turkiyyah	High Performance Tile Low Rank Symmetric Factorizations using Adaptive Randomized Approximation

Randomization techniques for solving linear systems of equations and eigenvalue problems
Laura Grigori

In this talk we discuss randomization techniques for solving large scale linear algebra problems. We focus in particular on solving linear systems of equations and eigenvalue problems and we present a randomized Gram-Schmidt process for orthogonalizing a set of vectors. We discuss its efficiency and its numerical stability while also using mixed precision. Its usage in the GMRES method for solving systems of equations is further presented. Application to block orthogonalization and eigenvalue problems is also addressed.

Adaptive Precision Solvers for Sparse and Data Sparse Systems
Theo Mary

The efficient solution of large sparse systems of linear equations often relies on various approximations to accelerate the computations, such as low-rank approximations or incomplete factorizations, whose quality and cost is controlled via a threshold parameter. The growing support of low precision floating-point arithmetics in hardware provides new opportunities and a new perspective. Rather than the current approach based on a single threshold, we should move towards a more gradual—almost continuous—use of approximations by adapting the precision to the data: store increasingly small elements in increasingly low precision. I will describe such adaptive precision solvers in the context of both direct and iterative sparse solvers, and showcase their potential on various real-life applications.

H2Pack: Software for H2 Hierarchical Matrices Using the Proxy Point Method
Hua Huang, Xin Xing, Edmond Chow

The H2 hierarchical matrix representation makes it possible to store certain dense kernel matrices and perform matrix-vector multiplications with them both in linear complexity with the number of points (rows of the kernel matrix). This makes it promising to use H2 hierarchical matrices for large-scale scientific and machine learning applications where these kernel matrices arise. H2Pack is a shared-memory parallel library for constructing and operating with H2 matrix representations for kernel matrices. It uses a new proxy point method for efficiently constructing the H2 matrix representation that works even for Gaussian kernel functions in machine learning. This talk reviews methods for efficiently constructing H2 hierarchical matrix representations (i.e., without forming the kernel matrix entirely) and compares and contrasts H2Pack with libraries for the fast multipole method.

High Performance Tile Low Rank Symmetric Factorizations using Adaptive Randomized Approximation
George Turkiyyah, Wajih Boukaram, Stefano Zampini, David Keyes

Tile low rank (TLR) representations of dense matrices partition them into blocks of roughly uniform size, where each off-diagonal tile is compressed and stored as its own low rank factorization. They offer an attractive representation for many data-sparse dense operators that appear in practical

applications, where substantial compression and a much smaller memory footprint can be achieved. TLR matrices are a compromise between the simplicity of a regular perfectly-strided data structure and the optimal complexity of the unbalanced trees of hierarchically low rank matrices, and provide a convenient performance-tuning parameter through their tile size that can be proportioned to take into account the cache size where the tiles reside in the memory hierarchy.

Despite their utility, however, there are currently no high performance algorithms that can generate their Cholesky and LDL^T factorizations and operate on them efficiently, particularly on GPUs. The difficulties in achieving high performance when factoring TLR matrices come from the expensive compression operations that must be performed during the factorization process and the adaptive rank distribution of the tiles that causes an irregular work pattern for the processing cores. In this work, we develop a dynamic batching operation and combine it with batched adaptive randomized approximations to remedy these difficulties and achieve high performance both on GPUs and CPUs.

Our implementation attains over 1.2 TFLOP/s in double precision on the V100 GPU, and is limited primarily by the underlying performance of batched GEMM operations. The time-to-solution also shows substantial speedup compared to regular dense factorizations. The Cholesky factorization of covariance matrix of size N = 131K arising in 2D or 3D spatial statistics, for example, can be factored to an accuracy $\varepsilon = 10^{-2}$ in just a few seconds. We believe the proposed GEMM-centric algorithm allows it to be readily ported to newer hardware such as the tensor cores that are optimized for small GEMM operations.

SESSION 8 — Low rank approximation, variable precision, randomization II

Esragul Korkmaz	Deciding Non-Compressible Blocks in Sparse Direct Solvers using Incomplete Factorization
Marek Felšöci	Direct solution of larger coupled sparse/dense FEM/BEM linear systems using low-rank compression
Bastien Vieublé	A mixed precision strategy for preconditioned GMRES

Deciding Non-Compressible Blocks in Sparse Direct Solvers using Incomplete Factorization

Esragul Korkmaz, Mathieu Faverge, Grégoire Pichon, Pierre Ramet

Low-rank compression techniques are very promising for reducing memory footprint and execution time on a large spectrum of linear solvers. Sparse direct supernodal approaches are one these techniques. However, despite providing a very good scalability and reducing the memory footprint, they suffer from an important flops overhead in their unstructured low-rank updates. As a consequence, the execution time is not improved as expected. In this work, we present a solution to improve low-rank compression techniques in sparse supernodal solvers. The proposed method tackles the overprice of the low-rank updates by identifying the blocks that have poor compression rates. We show that block incomplete LU factorization, thanks to the block fill-in levels, allows to identify most of these non-compressible blocks at low cost. This identification enables to postpone the low-rank compression step to trade small extra memory consumption for a better time to solution. The solution is validated within the PaStiX library with a large set of application matrices. It demonstrates sequential and multi-threaded speedup up to 8.5x, for small memory overhead of less than 1.49x with respect to the original version.

Direct solution of larger coupled sparse/dense FEM/BEM linear systems using low-rank compression

Emmanuel Agullo, Marek Felšöci, Guillaume Sylvand

In the aeronautical industry, aeroacoustics is used to model the propagation of acoustic waves in air flows enveloping an aircraft in flight. This for instance allows one to simulate the noise produced at ground level by an aircraft during the takeoff and landing phases, in order to validate that the regulatory environmental standards are met. Unlike most other complex physics simulations, the method resorts to solving coupled sparse/dense systems. Indeed, the heterogeneity of the jet flow created by reactors often requires a Finite Element Method (FEM) discretization, leading to a sparse linear system, while it may be reasonable to assume as homogeneous the rest of the space and hence model it with a Boundary Element Method (BEM) discretization, leading to a dense system. In an industrial context, these simulations are often operated on modern multicore workstations with fully-featured linear solvers. Exploiting their low-rank compression techniques is thus very appealing for solving larger coupled sparse/dense systems (hence ensuring a finer solution) on a given multicore workstation, and – of course – possibly do it fast. The standard method performing an efficient coupling of sparse and dense direct solvers is to rely on the Schur complement functionality of the sparse direct solver. However, to the best of our knowledge, modern fully-featured sparse direct solvers offering this functionality return the Schur complement as a non compressed matrix. We have studied the opportunity to process larger systems in spite of this constraint. For that we propose two classes of algorithms, namely multi-solve and multi-factorization, consisting in composing existing parallel sparse and dense methods on well chosen submatrices. An experimental study conducted on a 24 cores machine equipped with 128 GiB of RAM shows that these algorithms, implemented on top of state-of-the-art sparse and dense direct solvers, together with proper low-rank assembly schemes, can respectively process systems of 9 million and 2.5 million total unknowns instead of 1.3 million unknowns with a standard coupling of compressed sparse and dense solvers. We are currently extending this work to the out-of-core and distributed-memory cases. Moreover, we are working on a task-based implementation scheme allowing for a better inter-operability between the sparse and the dense solver and aiming at eliminating the current limitations of the multi-solve and multi-factorization schemes.

A mixed precision strategy for preconditioned GMRES

Patrick Amestoy, Alfredo Buttari, Nicholas J. Higham, Jean-Yves L'Excellent, Theo Mary, Bastien Vieublé

The new promises of accessible and efficient hardware support for very low precision arithmetics are a potential source for major performance improvements in scientific computing. However, exploiting such low arithmetic precisions while keeping a satisfactory accuracy on the solution leads to rethink our algorithms within a mixed precision setting. In this talk, we particularly focus on the use of mixed precision inside a left-preconditioned GMRES for the solution of linear systems. We especially cover a strategy consisting in applying the matrix-vector products with the original matrix A and the preconditioner in two different precisions. In particular, in some cases, the preconditioner can be applied in a lower precision than the matrix-vector product with A, leading to possible performance improvement when the application of the preconditioner is the dominant operation in an iteration of GMRES. We will demonstrate why and when this strategy makes sense by carrying a rounding error analysis on the algorithm and by providing numerical experiments using different preconditioners in Julia.

SESSION 9 — High performance computing III

Life with and after Fugaku — Exascale and Beyond
Satoshi Matsuoka

Fugaku is the first of the era of the so-called 'exascale' machine in which orders of magnitude improvement in performance has been attained compared to the 'petascale' supercomputers, serving as a platform of the most difficult societal challenges to satisfy SDGs. For this purpose Fugaku was designed not only to be extremely high-performant, but very general purpose with broad applicability and user base, allowing groundbreaking applications to be utilized, often converging traditional simulations, big data instrumentation and ML/AI, which will be introduced in the talk. Moreover, experiences with Fugaku is providing deep insights into how next generation machines should be designed, in particular the emphasis on optimizing data movement rather than achieving high flops count would be essential, or namely, "FLOPS to BYTES" being the most important technical aspect.

SESSION 10 — Least squares problems and optimization I

Jennifer Scott	Solving large linear least squares problems with equality constraints
Andy Wathen	Preconditioning for Normal Equations and Least Squares
Jemima Tabeart	Stein-based preconditioners for weak-constraint 4D-var
Nicolas Nadisic	Matrix-wise L0-constrained Sparse Nonnegative Least Squares

Solving large linear least squares problems with equality constraints
Jennifer Scott, Miroslav Tuma

Our interest is in solving large sparse linear least squares problems that are subject to one or more linear constraints that must be satisfied exactly. While some classical approaches are theoretically well founded, they can face difficulties when the matrix of constraints contains dense rows or if an algorithmic transformation used in the solution process results in a modified problem that is much denser than the original one. We propose a number of possible approaches with an emphasis on requiring that the constraints be satisfied with a small residual. Numerical experiments on problems coming from practical applications are used to demonstrate the effectiveness of the different ideas.

Preconditioning for Normal Equations and Least Squares
Andy Wathen

The solution of systems of linear(ized) equations lies at the heart of many problems in Scientific Computing. In particular for large systems, iterative methods are a primary approach. For many symmetric (or self-adjoint) systems, there are effective solution methods based on the Conjugate Gradient method (for definite problems) or minres (for indefinite problems) in combination with an appropriate preconditioner, which is required in almost all cases. For nonsymmetric systems there are two principal lines of attack: the use of a nonsymmetric iterative method such as gmres, or transformation into a symmetric problem via the normal equations. In either case, an appropriate preconditioner is generally required. We consider the possibilities here, particularly the idea of preconditioning the normal equations via approximations to the original nonsymmetric matrix. We highlight dangers that readily arise in this approach. Our comments also apply in the context of linear least squares problems as we will explain.

Stein-based preconditioners for weak-constraint 4D-var
Jemima Tabeart, Davide Palitta

The saddle point formulation of weak-constraint four-dimensional data assimilation offers the possibility of exploiting modern computer architectures and algorithms due to itsunderlying block structure. Developing good preconditioners which retain the highly-structured nature of the saddle point system has been an area of recent research interest, especially for applications to numerical weather prediction. In this talk I will present a new preconditioning approach which exploits inherent Kronecker structure within a matrixGMRES implementation. I will present theoretical results comparing our new preconditioners to existing standard choices of preconditioners. Finally I will present two numerical experiments for the heat equation and Lorenz 96 problem and show that our new approaches are competitive compared to current state-of-the-art preconditioners.

Matrix-wise L0-constrained Sparse Nonnegative Least Squares
Nicolas Nadisic, Jeremy Cohen, Arnaud Vandaele, Nicolas Gillis

Nonnegative least squares problems with multiple right-hand sides (MNNLS) arise in models that rely on additive linear combinations. In particular, they are at the core of most nonnegative matrix factorization algorithms and have many applications. The nonnegativity constraint is known to naturally favor sparsity, that is, solutions with few non-zero entries. However, it is often useful to further enhance this sparsity, as it improves the interpretability of the results and helps reducing noise, which leads to the sparse MNNLS problem. In this paper, as opposed to most previous works that enforce sparsity column- or row-wise, we first introduce a novel formulation for sparse MNNLS, with a matrix-wise sparsity constraint. Then, we present a two-step algorithm to tackle this problem. The first step divides sparse MNNLS in subproblems, one per column of the original problem. It then uses different algorithms to produce, either exactly or approximately, a Pareto front for each subproblem, that is, to produce a set of solutions representing different tradeoffs between reconstruction error and sparsity. The second step selects solutions among these Pareto fronts in order to build a sparsity-constrained matrix that minimizes the reconstruction error. We perform experiments on facial and hyperspectral images, and we show that our proposed two-step approach provides more accurate results than state-of-the-art sparse coding heuristics applied both column-wise and globally.

SESSION 11 — Least squares problems and optimization II

Alexandre Scotto Di Perrotolo	Towards efficient randomized limited memory preconditioners for variational data assimilation
Michal Kocvara	An interior-point method for Lasserre relaxations of unconstrained binary quadratic optimization problems
Mike Saunders	Algorithm NCL for constrained optimization

Towards efficient randomized limited memory preconditioners for variational data assimilation
Alexandre Scotto Di Perrotolo, Youssef Diouane, Selime Gurol, Xavier Vasseur

In variational data assimilation, a widespread method to solve the underlying optimization problem is the truncated Gauss-Newton method. In this method, each new descent direction is obtained as the solution of a potentially large-scale linear system involving a symmetric positive definite linear operator. A classical method to obtain an approximate solution is the Preconditioned Conjugate Gradient (PCG). The structure of variational data assimilation problems yields a natural preconditioner which is generally improved using a Limited Memory Preconditioner (LMP). This

so-called two-level preconditioner aims at accelerating the convergence of the method by integrating eigeninformation. The computation of exact eigeninformation being computationally out of reach for large-scale problems, standard strategies rather rely on Ritz approximations, computed from previous Krylov subspaces.

In this talk, we will present a class of randomized LMPs where the approximate eigenpairs are obtained using randomized procedures notably known to be well suited to large-scale computations. Unlike prior randomized methods, our approach is applicable when no factorization of the preconditioner is available. Here, we provide LMPs adapted for two variants of the PCG developed in the context of operational variational data assimilation, namely the inverse-free PCG and the augmented restricted PCG. The obtained algorithms can handle varying first-level preconditioners making them attractive regarding future designs of more elaborate preconditioning strategies. Numerical illustrations on a four-dimensional variational data assimilation problem will be presented, and demonstrate the potential of our approach.

An interior-point method for Lasserre relaxations of unconstrained binary quadratic optimization problems

Michal Kočvara, Soodeh Habibi, Michael Stingl

The aim of this talk is to solve linear semidefinite programs arising from Lasserre relaxations of unconstrained binary quadratic optimization problems. For this we use an interior point method with a preconditioned conjugate gradient method solving the linear systems. The preconditioner utilizes the low-rank structure of the solution of the relaxations. In order to fully utilize this, we re-write the moment relaxations. To treat the arising linear equality constraints we use an ℓ_1-penalty approach within the interior-point solver. The efficiency is demonstrated by numerical experiments and comparison with a state of the art semidefinite solver.

Algorithm NCL for constrained optimization

Michael Saunders

Algorithm NCL (nonlinearly constrained Lagrangian) was derived as a reimplementation of the LANCELOT augmented Lagrangian method. It shares LANCELOT's suitability for problems that do not satisfy the LICQ at a solution. About 10 nonlinearly constrained subproblems are solved by IPOPT or KNITRO, with warm starts on each subproblem. (These interior methods can be warmstarted!)

The original version of NCL was implemented in about 100 lines of AMPL to solve optimal taxation models with (for example) 500,000 nonlinear inequality constraints and 1000 variables plus bounds. The current Julia version is interfaced to the CUTEst test set. We consider Algorithm NCL's suitability as a general-purpose solver for problems where second derivatives are available.

Joint work with Ding Ma and Dominique Orban.

POSTER BLITZ

Théo Beuzeville	Adversarial attacks via sequential quadratic programming
Andrei Dumitrasc	Inexact inner-outer Golub-Kahan bidiagonalization method: a relaxation strategy
Quentin Ferro	Neural Network Precision Tuning Using Stochastic Arithmetic
Matthieu Gerest	Mixed precision block low-rank compression for the solution of sparse linear systems
Sadok Jerad	Optimal second-order complexity without function evaluations
Sophie Mauran	Introduction of kernel methods in data assimilation
Roméo Molina	Adaptive Precision Sparse Matrix-Vector Product and its Application to Krylov Solvers
Daichi Mukunoki	Remedies for Reproducibility Issue in Conjugate Gradient Solvers
Mathis Peyron	Latent space data assimilation by using deep learning

Adversarial attacks via sequential quadratic programming
Théo Beuzeville, Alfredo Buttari, Serge Gratton, Theo Mary, Erkan Ulker, Pierre Boudier

Sequential Quadratic Programming (SQP) is an iterative method, one of the most successful for constrained nonlinear optimization. It provides powerful algorithmic tools for the solution of large-scale problems by solving a sequence of constrained quadratic problems.

While deep neural networks (DNN) have achieved an unprecedented success in numerous machine learning tasks in various domains, their robustness to adversarial attacks, rounding errors, or quantization processes has raised considerable concerns from the machine learning community.

In this work we propose a novel approach for the construction of adversarial attacks which relies on a local SQP strategy. These attacks, using second order information produce smaller perturbations, in norm, than perturbations obtained with existing gradient-based approaches. We produce numerical results that support our theoretical findings and illustrate the relevance of our approach on well-known datasets.

Inexact inner-outer Golub-Kahan bidiagonalization method: a relaxation strategy
Vincent Darrigrand, Andrei Dumitrasc, Carola Kruse, Ulrich Rüde

We consider the generalized Golub-Kahan bidiagonalization in the context of inner-outer Krylov schemes. The inner linear system is solved iteratively, introducing a perturbation. By controlling the magnitude of its norm, we can obtain an outer solution of desired accuracy and reduce the cost of the inner solver. This is achieved by dynamically changing the inner tolerance, reducing the number of iterations performed. As test cases, we take the Stokes flow problem and a mixed formulation of the Poisson problem.

Neural Network Precision Tuning Using Stochastic Arithmetic
Quentin Ferro, Stef Graillat, Thibault Hilaire, Basile Lewandowski, Fabienne Jézéquel

Neural networks can be costly in terms of memory and computing. Reducing their cost has become an objective, especially when integrated in an embedded system with limited resources. A solution is often to reduce the precision of their neurons parameters. In this article, we present how to use auto-tuning on neural networks to lower their precision while keeping an accurate output. To do so, we use a floating-point auto-tuning tool on different kinds of neural networks. We show that, to some extent, we can lower the precision of several neural network parameters without compromising the accuracy requirement.

Mixed precision block low-rank compression for the solution of sparse linear systems
Patrick Amestoy, Olivier Boiteau, Alfredo Buttari, Matthieu Gerest, Fabienne Jézéquel, Jean-Yves L'Excellent, Theo Mary

Many applications involve data-sparse matrices whose off-diagonal blocks have low numerical ranks. Block low-rank (BLR) compression allows us to exploit this property, by representing those blocks as a truncated SVD or QR decomposition. Thus the costs for storing the matrix and computing its LU factorization can both be reduced, while controlling the overall error via a truncation parameter ε.

We present a mixed precision variant of BLR compression, that uses several precision formats simultaneously. We prove that most coefficients and most operations can be safely switched from double to lower precision formats, such as single or even half precision. We apply this approach to the solution of large scale linear systems with the multifrontal solver MUMPS. We obtain significant storage reductions for a range of real-life industrial applications.

Optimal second-order complexity without function evaluations
Sadok Jerad, Serge Gratton, Philippe L. Toint

An adaptive regularization algorithm for unconstrained nonconvex optimization is presented in which the objective function is never evaluated, but only derivatives are used. This algorithm belongs to the class of adaptive regularization methods, for which optimal worst-case complexity results are known for the standard framework where the objective function is evaluated. It is shown in this presentation that these excellent complexity bounds are also valid for the new algorithm, despite the fact that significantly less information is used. The new algorithm using first and second derivatives, when applied to functions with Lipschitz continuous Hessian, will find an iterate at which the gradient's norm with the same complexity as standard adaptive regularization algorithm that uses the function values.

Introduction of kernel methods in data assimilation
Sophie Mauran, Ehouarn Simon, Sandrine Mouysset, Laurent Bertino

Data assimilation methods, the most commonly used ones from an operational forecasting point of view, can be interpreted as a Bayesian estimation process, under Gaussian assumptions on the errors and variables. However, these assumptions are not always valid, which can lead to their divergence. The work presented here proposes the use of kernel methods to extend these methods to problems for which their performance remains limited. We propose a formulation of the Ensemble transform Kalman filter (ETKF) from the point of view of an optimisation problem on the Reproducing Kernel Hilbert Space (RKHS) associated to the linear kernel, and a generalisation of this approach for any RKHS. We propose a new formulation of this algorithm, extended to any kernel, and evaluate the performance of this approach on a toy model representative of the dynamics of a geophysical fluid, for the case of linear, polynomial and Gaussian kernels.

Adaptive Precision Sparse Matrix-Vector Product and its Application to Krylov Solvers
Roméo Molina

We introduce a mixed precision algorithm for computing sparse matrix-vector products and use it to accelerate the solution of sparse linear systems by iterative methods. Our approach is based on the idea of adapting the precision of each matrix element to their magnitude: we split the elements into buckets and use progressively lower precisions for the buckets of progressively smaller elements. We carry out a rounding error analysis of this algorithm that provides us with an explicit rule to decide which element goes into which bucket and allows us to rigorously control the accuracy of the algorithm. We implement the algorithm on a multicore computer and obtain significant speedups (up to a factor 7.5) with respect to uniform precision algorithms, without loss of accuracy, on a range of

sparse matrices from real-life applications. We showcase the effectiveness of our algorithm by plugging it into a GMRES solver for sparse linear systems and observe that the convergence of the solution is essentially unaffected by the use of adaptive precision.

Remedies for Reproducibility Issue in Conjugate Gradient Solvers

Daichi Mukunoki, Roman Iakymchuk, Fabienne JEZEQUEL, Katsuhisa Ozaki, Takeshi Ogita, Toshiyuki Imamura

In parallel computing, sparse iterative solvers often fail to reproduce a result due to the effect of rounding errors. For example, slightly different results with different convergence histories may be observed on different CPUs and GPUs. This can be a problem for debugging and porting codes to different systems, as well as an obstacle for quality assurance and discussions in scientific activities. In this poster, we present several remedies for the reproducibility issue in Conjugate Gradient methods on many-core processors. We show the performance and results obtained using three different approaches: (1) ensuring 100% bit-wise reproducibility using the infinite-precision BLAS, (2) enhancing the possibility of reproducibility using the high-precision BLAS, and (3) evaluating the reproducibility probabilistically using numerical validation.

Latent space data assimilation by using deep learning

Mathis Peyron

Performing Data Assimilation at a low cost is of prime concern in Earth system modeling, particularly at the time of big data where huge quantities of observations are available. Capitalizing on the ability of Neural Networks techniques for approximating the solution of PDE's, we incorporate Deep Learning methods into a DA framework. More precisely, we exploit the latent structure provided by autoencoders to design an Ensemble Transform Kalman Filter with model error (ETKF-Q) in the latent space. Model dynamics are also propagated within the latent space via a surrogate neural network.

This novel ETKF-Q-Latent algorithm is tested on a tailored instructional version of Lorenz 96 equations, named the augmented Lorenz 96 system: it possesses a latent structure that accurately represents the observed dynamics. Numerical experiments based on this particular system evidence that the ETKF-Q-L approach both reduces the computational cost and provides better accuracy than state of the art algorithms, such as the ETKF-Q.