Sparse Days

September 6th – 8th, 2017

CERFACS, Toulouse, France

CERFACS
CENTRE EUROPÉEN DE RECHERCHE ET DE FORMATION AVANCÉE EN CALCUL SCIENTIFIQUE

IRIT
CNRS - INPT - UPS - UT1 - UT2J

cimi
Centre International
de Mathématiques et d'Informatique
TOULOUSE

INVESTISSEMENTS D'AVENIR

Sparse Days programme
Wednesday, September 6th 2017

13:00 – 14:00 Registration
14:00 – 14:15 Welcome message

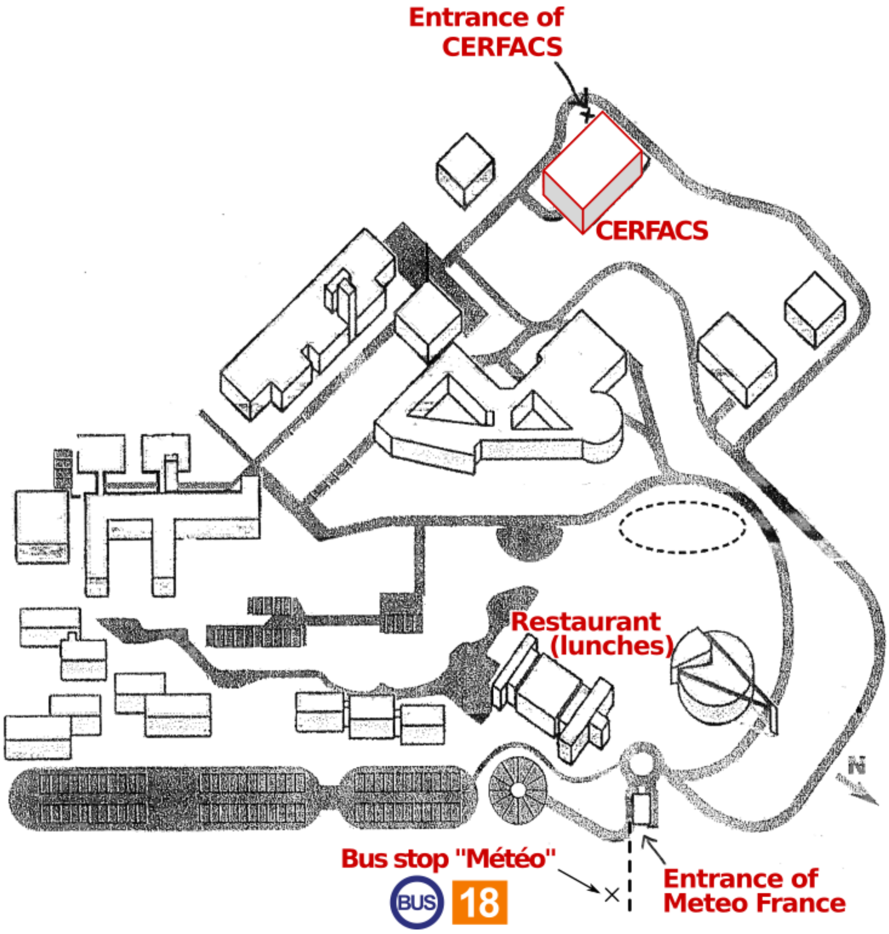
Session 1 (14:15 – 15:30) – Chair: Iain Duff

14:15 – 15:40 Block Low-Rank multifrontal solvers: complexity, performance, and scalability
Theo Mary (IRIT)
15:40 – 16:05 Sparse Supernodal Solver exploiting Low-Rankness Property
Grégoire Pichon (Inria Bordeaux – Sud-Ouest)
15:05 – 15:30 Recursive inverse factorization
Anton Artemov (Uppsala University)

Coffee break (15:30 – 16:00)

Session 2 (16:00 – 17:40) – Chair: Daniel Ruiz

16:00 – 16:25 *Algorithm based on spectral analysis to detect numerical blocks in matrices*
Luce le Gorrec (IRIT – APO)
16:25 – 16:50 Computing Sparse Tensor Decompositions using Dimension Trees
Oguz Kaya (ENS Lyon)
16:50 – 17:15 Numerical analysis of dynamic centrality
Philip Knight (University of Strathclyde)
17:15 – 17:40 SuiteSparse:GraphBLAS: graph algorithms via sparse matrix operations on semirings
Tim Davis (Texas A&M University)



Thursday, September 7th 2017

Session 3 (09:45 – 11:00) – Chair: Tim Davis

09:45 – 10:10 Parallel Algorithm Design via Approximation
Alex Pothen (Purdue University)
10:10 – 10:35 Accelerating the Mondriaan sparse matrix partitioner
Rob Bisseling (Utrecht University)
10:35 – 11:00 LS-GPart: a global, distributed ordering library
François-Henry Rouet (Livermore Software Technology Corporation)

Coffee break (11:00 – 11:25)

Session 4 (11:25 – 13:05) – Chair: Cleve Ashcraft

11:25 – 11:50 A Tale of Two Codes
Robert Lucas (Livermore Software Technology Corporation)
11:50 – 12:15 Direct solution of sparse systems of linear equations with sparse multiple right hand sides
Gilles Moreau (ENS Lyon)
12:15 – 12:40 Averaged information splitting and multi-frontal solvers for heterogeneous high-throughput data analysis
Shengxin Zhu (Xi'an Jiaotong – Liverpool University)
12:40 – 13:05 Permuting Spiked Matrices to Triangular Form in the Linear Programming LU Update
Lukas Schork (University of Edinburgh, UK)

Lunch break (13:05 – 15:00)

Session 5 (15:00 – 16:15) – Chair: Patrick Amestoy

15:00 – 15:25 Factorization Based Sparse Solvers and Preconditioners for Exascale
Sherry Li (Lawrence Berkeley National Laboratory)
15:25 – 15:50 Scalability Analysis of Sparse Matrix Computations on Many-core Processors
Weifeng Liu (Norwegian University of Science and Technology)
15:50 – 16:15 Refactoring Sparse Triangular Solver on Sunway TaihuLight Many-core Supercomputer
Wei Xue (Tsinghua University)

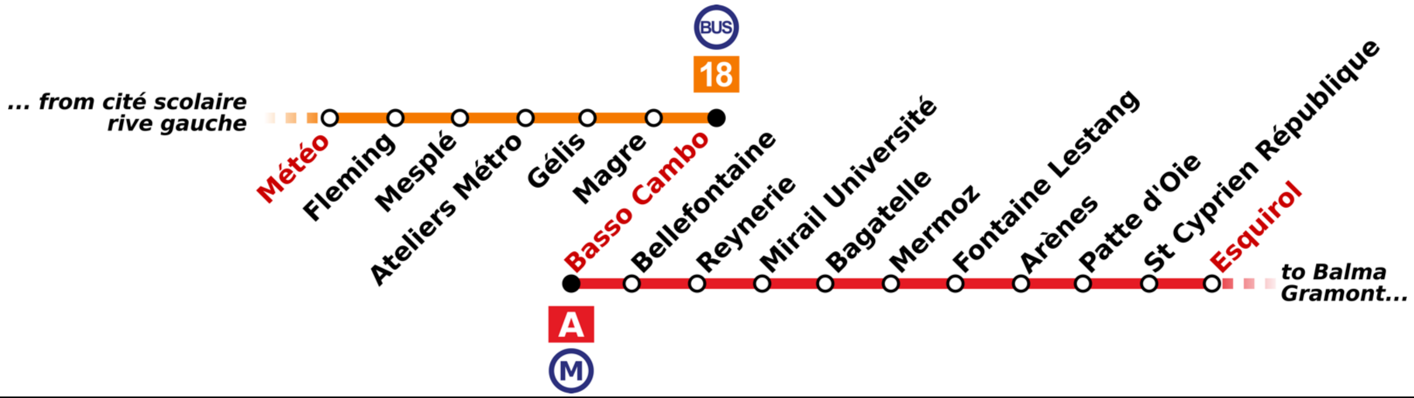
Coffee break (16:15 – 16:40)

16:40 – 17:20 Celebration of Iain Duff's birthday.
17:20 – 17:40 Review of Sparse Days by Iain Duff.

Reception dinner (19:45) – Brasserie Flo Les Beaux-Arts, 1 quai de la Daurade, Toulouse

Reception dinner – directions from Météo-France to metro station Esquirol:

Get on **bus 18 towards "Basso Cambo"** at stop "Météo" (close to the entrance of Météo France).
Get off at **stop "Basso Cambo"** (end of the line).
Take **metro A** (towards "Balma Gramont") and get off at **station "Esquirol"**.
Walk (5 min) from station "Esquirol" to the restaurant (see plan on next page).



Friday, September 8th 2017

Session 6 (09:45 – 11:00) – Chair: Åke Björck

09:45 – 10:35 On solving mixed sparse-dense linear least-squares problems
Part I: A Schur complement approach
Part II: A preconditioned conjugate gradient method
Jennifer Scott (STFC Rutherford Appleton Laboratory); and
Miroslav Tuma (Charles University and Czech Academy of Sciences)

10:35 – 11:00 *A Parallel Hierarchical Low-Rank Solver for General Sparse Matrices*
Erik Boman (Sandia National Laboratories)

Coffee break (11:00 – 11:20)

Session 7 (11:20 – 13:00) – Chair: Daniel Loghin

11:20 – 11:45 High Performance Block Incomplete LU Factorization
Matthias Bollhoefer (TU Braunschweig)

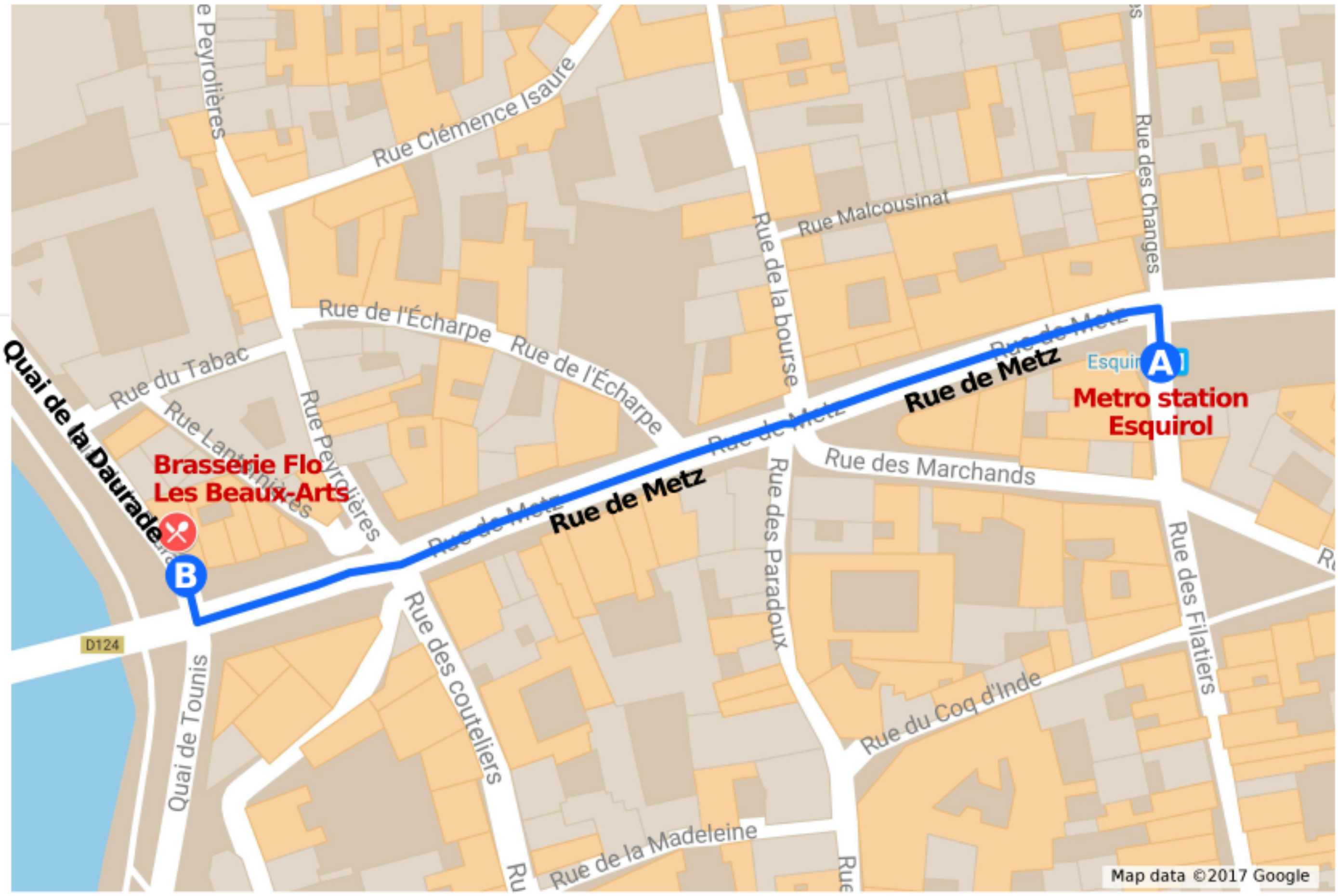
11:45 – 12:10 *An optimal Q-OR method for solving nonsymmetric linear systems*
Gérard Meurant

12:10 – 12:35 *Preconditioning for non-symmetric Toeplitz matrices with application to time-dependent PDEs*
Andrew Wathen (Oxford University)

12:35 – 13:00 Efficient Parallel iterative solvers for high order DG methods
Matthias Maischak (Brunel University)

Lunch (13:00)

Reception dinner – directions from metro station Esquirol (A) to Brasserie Flo Les Beaux-Arts (B):



Abstracts

Recursive inverse factorization

Anton Artemov (Uppsala University, Sweden)

In this talk we will present a parallel algorithm for the inverse factorization of symmetric positive definite matrices. This algorithm utilizes a hierarchical representation of the matrix and computes an inverse factor by combining an iterative refinement procedure with a recursive binary principal submatrix decomposition. Although our algorithm should be generally applicable, this work was mainly motivated by the need for efficient and scalable inverse factorization of the basis set overlap matrix in large scale electronic structure calculations. We show that for such matrices the computational cost increases only linearly with system size. We will also demonstrate the weak scaling behavior of the algorithm and give theoretical justification of the scaling behavior based on critical path length estimation.

Accelerating the Mondriaan sparse matrix partitioner

Rob Bisseling (Mathematical Institute, Utrecht University, Netherlands)

The Mondriaan package for sparse matrix partitioning has been developed since 2002 and currently it has reached version 4.1. Its latest addition is the medium-grain method for 2D partitioning, which significantly reduces the communication volume of a parallel sparse matrix-vector multiplication. The present talk is about recent developments to be included in version 4.2 of Mondriaan, which focus on improving the speed of computing the partitioning. One improvement is the use of free nonzeros in the matrix, which can be moved after a split to improve the load balance without adversely affecting the communication volume. This leaves more room for reducing communication in subsequent splits. Another improvement for certain sparse matrices is obtained by trying to find connected components after every split that satisfy the imposed imbalance constraint. If this succeeds, a zero-volume split is found very quickly. The total savings in computation time of these and other improvement for a set of 900 sparse matrices is about 25 per cent. This is joint work with Marco van Oort.

High Performance Block Incomplete LU Factorization

Matthias Bollhoefer (TU Braunschweig, Germany)

Many application problems that lead to solving linear systems make use of preconditioned Krylov subspace solvers to compute their solution.

Among the most popular preconditioning approaches are incomplete factorization methods either as single--level approaches or within a multilevel framework. We will present a block incomplete triangular factorization that is based on skillfully blocking the system initially and throughout the factorization. This approach allows for the use of cache-optimized dense matrix kernels such level-3 BLAS or LAPACK. We will demonstrate how this block approach may significantly outperform the scalar method on modern architectures, paving the way for its prospective use inside various multilevel incomplete factorization approaches or other applications where the core part relies on an incomplete factorization.

A Parallel Hierarchical Low-Rank Solver for General Sparse Matrices

Erik Boman (Sandia National Laboratories, USA)

Hierarchical methods that exploit low-rank structure have attracted much attention lately. For sparse matrices, many methods rely on low-rank structure within dense frontal matrices. We present a different approach by Eric Darve et al. that works on the entire sparse matrix, and is in several ways simpler. The method is most useful as a preconditioner, though it could also be used as a direct solver. We show that the method can be viewed as an H2 solver and also as an extension to the block ILU(0) preconditioner. We briefly discuss a parallel code developed in collaboration with Stanford, and show some results. Joint work with Chao Chen, Eric Darve, Siva Rajamanickam, and Ray Tuminaro.

SuiteSparse:GraphBLAS: graph algorithms via sparse matrix operations on semirings

Tim Davis (Texas A&M University, USA)

SuiteSparse:GraphBLAS is an 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.

Performance of SuiteSparse:GraphBLAS is either on par with the corresponding operations in MATLAB, or faster. Submatrix assignment is particularly efficient. In one example, C(I,J)=A for a matrix C of size 3 million-by- 3 million with 14 million nonzeros, and a matrix A of size 5500-by-7000 with 38500 nonzeros, takes 82 seconds in MATLAB but only 0.74 seconds in SuiteSparse:GraphBLAS. This result includes finalizing the computation and returning the result to MATLAB as a valid sparse matrix. SuiteSparse:GraphBLAS also includes a non-blocking mode, so that a sequence of submatrix assignments can be still more efficient.

Computing Sparse Tensor Decompositions using Dimension Trees

Oguz Kaya (ENS Lyon, France)

Tensor decompositions have increasingly been used in many applications domains including computer vision, signal processing, graph analytics, healthcare data analysis, recommender systems, and many other related machine learning problems.

In many of these applications, the tensor representation of data is big, sparse, and of low-rank, and an efficient computation of tensor decompositions is indispensible to be able to analyze datasets of massive scale.

Two major computational kernels involved at the core of the iterative tensor decomposition algorithms are tensor-times-vector (TTV) and -matrix (TTM) multiplications.

In this talk, we propose a novel algorithmic framework along with an associated sparse tensor data structure that enables an efficient computation of these two operations for high dimensional sparse tensors.

With this framework, we enable storing and reusing partial TTV and TTM results, thereby reducing the cost per iteration to O(NlogN) TTM/TTV calls from O(N^2) in the traditional scheme.

Second, we propose an efficient tensor storage scheme representing an N-dimensional tensor using O(NlogN) index arrays, and enabling the partial computations in this framework.

Third, we propose shared memory parallelization of TTV and TTMs using this sparse data representation.

With all these optimizations, we up to 6x speedup over the state of the art implementations using 32 dimensional sparse tensors.

Numerical analysis of dynamic centrality

Philip Knight (University of Strathclyde, UK)

Centrality measures have proven to be a vital tool for analysing static networks. In recent years, many of these measures have been adapted for use on dynamic networks. In particular, spectral measures such as Katz centrality have been extended to take into account time's arrow. We show that great care must be taken in applying such measures due to the inherent ill-conditioning in the associated matrices (an ill-conditioning which can get so extreme that the matrices involved can have a numerical rank of 1). At the same time, the ill-conditioning can reduce computational effort. We investigate some pre-conditioning techniques to alleviate the ill-conditioning, while attempting to keep the costs of computation low.

Algorithm based on spectral analysis to detect numerical blocks in matrices

Luce le Gorrec (IRIT-APO, France)

Block detection in matrices is involved in several current applied mathematics fields, such as block preconditioning perspectives in the context of sparse linear systems [1], [2], community detection in networks [3], and data mining – for instance in biological applications [4] - which requires to consider the case of dense and sparse matrices.

We are setting up an algorithm with as general a design as possible, making it able to deal with various kinds of matrices and different applications.

As we wish to address both sparse and dense matrix cases, we exploit together structural and numerical information. Structural information is analysed via classical tools from graph partitioning, and to highlight numerical substructuring we apply a preliminary doubly stochastic scaling stage. The numerical blocks are then expected to be revealed throughout spectral elements of such a scaled matrix.

Although spectral clustering is not something new [5], we base our cluster analysis on the numerical structure of singular vectors only, mostly by means of tools from signal processing.

This approach is still consistent whatever the application, which is not the case for the spectral methods based on clustering measures as modularity or conductance [6]. Moreover, those algorithms can also depend on some undesired specific behavior of such measures. In our case, the clustering obtained because of the singular vectors depends on the scaled matrix only.

However, we also exploit these measures a posteriori in a validation stage, because they are useful to quantify the resulting clustering. This last validation stage is therefore application dependent.

We finally discuss the effectiveness of our algorithm with results on some well-known networks.

References:

[1] Eugene Vecharynski, Yousef Saad, and Masha Sosonkina, Graph partitioning using matrix values for preconditioning symmetric positive definite systems. SIAM Journal on Scientific Computing, 36(1):A63–A87, 2014

[2] Yao Zhu and Ahmed H. Sameh, How to Generate Effective Block Jacobi Preconditioners for Solving Large Sparse Linear Systems, pages 231–244. Springer International Publishing, Cham, 2016.

[3] P. Conde-Céspedes, Modélisation et extension du formalisme de l'analyse relationnelle mathématique à la modularisation des grands graphes. Phd thesis, Université Pierre et Marie Curie, Paris, France, Chap 3, 2013

[4] V. Calderon, R. Barriot, Y. Quentin, and G. Fichant, Crossing isorthology and microsynteny to resolve multigenic families functional annotation. 22nd International Workshop on Database and Expert Systems Applications, pages 440–444, 2011.

[5] F.R.K. Chung, Spectral Graph Theory. Number n 92 in CBMS Regional Conference Series. Conference Board of the Mathematical Sciences.

[6] Satu Elisa Schaeffer. Survey. Graph clustering. Comput. Sci. Rev., 1(1) :27–64, August 2007.

Factorization Based Sparse Solvers and Preconditioners for Exascale

Sherry Li (Lawrence Berkeley National Laboratory, USA)

Factorization based algorithms are often the most robust algorithmic choices for solving linear systems from multiphysics and multiscale simulations, being used as direct solvers, or as coarse-grid solvers in multigrid, or as preconditioners for iterative solvers.

As architectures become increasingly diverse, the challenges of developing high performance codes are becoming more onerous. We present our recent research on developing novel factorization algorithms with lower arithmetic complexity as well as lower communication requirement.

We will illustrate how to implement these algorthms on the architectures that are likely the building blocks of future exascale computers, including manycore nodes Xeon Phi Knights Landing and GPU, and demonstrate their performance on the latest manycore machines.

Scalability Analysis of Sparse Matrix Computations on Many-core Processors

Weifeng Liu (Norwegian University of Science and Technology, Norway)

Sparse matrices exist in a number of computational problems in scientific and engineering. Researchers have been always looking for faster parallel algorithms for sparse matrix in the last decades. Recently, the emergence of many-core processors has introduced more conflicts between algorithm efficiency and scalability. On one hand, many-core processors require a large amount of fine-grained tasks to saturate their resources, on the other, the irregular structure of sparse matrix brings difficulties to the task partitioning. This talk will discuss scalability of existing work and our newly designed data structures, such as the CSR5 format (ICS '15), and algorithms, such as sparse matrix transposition (ICS '16), sparse matrix-vector multiplication (ICS '15 and Parco), sparse triangular solve (Euro-Par '16) and sparse matrix-matrix multiplication (IPDPS '14 and JPDC). Several key challenges in this area will be presented as well.

A Tale of Two Codes

Robert Lucas (Livermore Software Technology Corportion, USA)

Since its invention by Duff and Reid, the multifrontal method has been realized in many different software implementations. There are many design design that the developers face, and they evolve as new linear systems and computer architectures are encountered. This talk will compare and contrast two such codes, the open source MUMPS, and a proprietary code used in LS-DYNA. We will highlight some differences made in the design of the two codes, and discuss their impact as systems have evolved.

This work is in collaboration with Patrick Amestoy (ENSEEIHT), Roger Grimes (LSTC), Jean-Yves L'Excellent (ENS Lyon), Theo Mary (ENSEEIHT), Chiara Pugliasi (MUMPS consortium), Francois-Henry Rouet (LSTC), and Clement Weisbecker (LSTC).

Efficient Parallel iterative solvers for high order DG methods

Matthias Maischak (Brunel University, UK)

We analyze how the efficient parallelization of iterative solvers for high order DG methods depends on the choice of basis functions. We describe a type of basis function which minimises communication costs for the matrix vector multiplication and does not significantly increase the condition number compared with standard choices. In addition we describe a data model which allows the parallelization of matrix vector multiplication and additive Schwarz preconditioner with minimal communication costs in 2d and 3d settings.

Block Low-Rank multifrontal solvers: complexity, performance, and scalability

Theo Mary (Université de Toulouse, IRIT, France)

We consider the use of Block Low-Rank format (BLR) to solve large sparse real-life problems with multifrontal direct solvers. In this talk, we present the different BLR factorization variants, depending on the strategies used to perform the updates in the frontal matrices and on the approaches to handle numerical pivoting. We compare these variants in terms of complexity, robustness, performance, and scalability. In our numerical experiments, the MUMPS solver is used to compare and analyze each BLR variant in a parallel (MPI+OpenMP) setting on a variety of applications.

An optimal Q-OR method for solving nonsymmetric linear systems

Gérard Meurant

Eiermann and Ernst showed that most Krylov methods for solving linear systems with nonsymmetric matrices can be described as so-called quasi-orthogonal (Q-OR) or quasi-minimum (Q-MR) residual methods. There exist many pairs of Q-OR/Q-MR methods. Well-known examples are FOM/GMRES, BiCG/QMR and Hessenberg/CMRH. These pairs mainly differ by the different bases of the Krylov subspace they used.

In this lecture we will first recall the generic properties of the Q-OR methods. Then, we will show how to construct a non-orthogonal basis of the Krylov subspace for which the Q-OR method yields the same residual norms as GMRES up to the final stagnation phase. Therefore, for a given Krylov subspace, this is the optimal Q-OR method for the residual norms. We will also established some properties of this new basis that will help us simplifying the implementation of the proposed algorithm.

We will illustrate the performances of the new algorithm with numerical experiments. In particular, for many linear systems, this new method gives a better attainable accuracy than GMRES using the modified Gram-Schmidt algorithm as well as GMRES using Householder reflections.

Direct solution of sparse systems of linear equations with sparse multiple right hand sides

Gilles Moreau (UCBL, ENS Lyon, France)

The cost of the solution phase of sparse direct solvers is sometimes critical. It can be larger than the cost of the factorization in applications where systems of linear equations with thousands of right-hand sides must be solved. Depending on the applications, the right-hand sides may be known all at once or may depend on previous solutions, and they may be dense or sparse. In this talk, we consider applications associated with 2D or 3D problems coming from electromagnetism or seismic modeling applications, which usually process multiple sparse right-hand sides with different non- zero structure represented by a matrix B. In direct methods, systems of linear equations of the form AX = B are then solved by first factorizing matrix (A = LU or LDL T), then by performing the solution phase (where LY = B and U X = Y are solved), based on dependencies given by a so called elimination tree. Exploiting sparsity in matrix B during the solve phase may lead to significant gains on the total number of operations. This may be done by (i) pruning nodes in the tree corresponding to zero entries in columns of B (vertical sparsity), and (ii) by working on subsets of columns at each node of the tree (horizontal sparsity). In this study, we explain why horizontal sparsity leads to the resolution of a combinatorial problem based on the permutation of columns of right hand sides. Then, we develop the main contribution which consists of two algorithms aiming at reducing the number of operations during the forward elimination (LY = B). The first one computes a suitable permutation of the columns of B. The second goes further by splitting columns of B in several groups, while creating new groups only when needed to preserve BLAS 3 effects and flexibility of the implementation. Both algorithms have geometrical motivations but an algebraic implementation, allowing them to work on arbitrary trees and matrices of right-hand sides. We present the numbers of operations reached by the proposed algorithms compared to classical techniques on our target applications, and show some preliminary performance results. Finally, we give some perspectives for this work.

Sparse Supernodal Solver exploiting Low-Rankness Property

Grégoire Pichon (Inria Bordeaux - Sud-Ouest, France)

In this talk, we will present recent advances on PaStiX, a supernodal sparse direct solver, which has been enhanced by the introduction of Block Low-Rank compression. We will describe different strategies leading to memory consumption gain and/or time-to-solution reduction. Finally, the implementation on top of runtime systems (Parsec, StarPU), will be compared with the static scheduling used in previous experiments.

This work is a co-joint work with Eric Darve, Mathieu Faverge, Pierre Ramet and Jean Roman

Parallel Algorithm Design via Approximation

Alex Pothen (Purdue University, USA)

We describe a paradigm for designing parallel algorithms via approximation techniques. Instead of solving a problem exactly, we seek a solution with provable approximation guarantees via the design of approximation algorithms that have high degrees of parallelism. We use matchings and edge covers in graphs as examples of this paradigm. We describe 1/2-approximation algorithm for maximum weighted b-matching, and 3/2- and 2-approximation algorithms for minimum weighted b-edge covers. We show that this technique leads to fast and scalable parallel algorithms for these problems on shared memory and distributed memory computers. We will also describe a few applications of matchings and edge covers.

LS-GPart: a global, distributed ordering library

François-Henry Rouet (Livermore Software Technology Corporation, USA)

Computing a fill-reducing ordering of a sparse matrix is a critical step for sparse direct solvers. Top-down methods such as Nested Dissection are often favored, in particular for problems arising from the discretization of PDEs. Most parallel implementations rely on the multilevel framework, in which the adjacency graph of the matrix is coarsened in order to find separators. The approach we propose here is a top-down algorithm, but it is not multilevel. Instead, we perform all the operations on the whole, uncoarsened, graph. We refer to this as a "global" method. We build vertex separators of the graph using "half-level sets", an extension of level sets used in the Cuthill-McKee algorithm and in George and Liu's automatic Nested Dissection. We present a distributed-memory implementation, that we compare against METIS and Scotch for problems arising from our multiphysics code LS-Dyna. We also investigate properties of our ordering in the context of Block Low-Rank factorizations.

Permuting Spiked Matrices to Triangular Form in the Linear Programming LU Update

Lukas Schork (University of Edinburgh, UK)

Commonly used methods for updating a sparse LU factorization after a column modification proceed in two steps: They first insert a "spike" column into U, and then apply algebraic operations to restore triangularity. In the linear programming basis update it frequently happens that the spiked matrix is already permuted triangular, i.e. it can be made upper triangular by a row and column permutation. This talk presents an efficient method to find these permutations, if they exist, and to combine them with the Forrest-Tomlin (FT) update. Numerical comparisons with the pure FT update on large scale linear programming problems are given.

On solving mixed sparse-dense linear least-squares problems
Part I: A Schur complement approach
Part II: A preconditioned conjugate gradient method

Jennifer Scott (STFC Rutherford Appleton Laboratory, UK); and
Miroslav Tuma (Charles University and Czech Academy of Sciences, Czech Republic)

The efficient solution of large linear least-squares problems in which the system matrix A contains rows with a large number of entries is challenging. In these talks, we present two very different approaches to tackling such problems.

Part I:
The first approach is based on the Schur complement method.
This reduces the problem to factorizing a reduced normal matrix $\mathbf{C}_s$ and then forming and factorizing a small dense Schur complement matrix; the solution process is completed by using the sparse and dense factors to solve a number of triangular systems. We look at using both complete and incomplete factorizations of $\mathbf{C}_s$ and discuss dealing with rank deficiency in $\mathbf{C}_s$.
Part II:
The second approach processes relatively dense rows of A within a conjugate gradient method with an incomplete factorization preconditioner together with the complete Cholesky factorization of a dense matrix of order equal to the number of rows identified as dense.

In both talks, numerical experiments on large-scale problems arising from practical applications are used to illustrate the effectiveness of the proposed approaches.
The results demonstrate that we are able to solve problems that are intractable if the dense rows are not exploited.

Preconditioning for non-symmetric Toeplitz matrices with application to time-dependent PDEs

Andrew Wathen (Oxford University, UK)

Gil Strang proposed the use of circulant matrices (and the FFT) for preconditioning symmetric Toeplitz (constant-diagonal) matrix systems in 1986 and there is now a well-developed theory which guarantees rapid convergence of the conjugate gradient method for such preconditioned positive definite symmetric systems.
In this talk we describe our recent approach which provides a preconditioned MINRES method with the same guarantees for real nonsymmetric Toeplitz systems regardless of the non-normality. We demonstrate the utility of these ideas in the context of time-dependent PDEs.
This is joint work with Elle McDonald and Jen Pestana.

Refactoring Sparse Triangular Solver on Sunway TaihuLight Many-core Supercomputer

Wei Xue (Tsinghua University, China)

The sparse triangular solver, SpTRSV, is one of the most important kernels in many scientific and engineering applications. Efficiently parallelizing the SpTRSV on modern manycore architectures is considerably difficult due to the inherent dependency of tasks, and the frequent but discontinuous memory accesses. Achieving high performance of SpTRSV is even more challenging for SW26010, the new-generation customized heterogeneous many-core processor equipped in the Sunway TaihuLight supercomputer. The known parallel SpTRSVs have to be refactored to fit the single-thread and cacheless design of SW26010. In this work, we focus on how to design and implement fast SpTRSV for sparse problems on SW26010. A generalized algorithm framework of parallel SpTRSV is proposed for best utilization of the features and flexibilities of SW26010 many-core architecture according to the fine-grained Producer-Consumer model. Moreover, a novel parallel SpTRSV is presented by using direct data transfers across registers of the computing elements of SW26010. Experiments on four typical structured-grid triangular problems with different problem sizes demonstrate that our SpTRSV can achieve an average bandwidth utilization of 81.7%, which leads to a speedup of 17 over serial method on the management processing element of SW26010. And experiments with structured-grid linear sparse solvers show that this new SpTRSV can achieve superior performance over the latest Intel Xeon CPU and Intel KNL over DDR4 memory.

Averaged information splitting and multi-frontal solvers for heterogeneous high-throughput data analysis

Shengxin Zhu (Xi'an Jiaotong-Liverpool University, China)

Linear mixed models are frequently used for analyzing heterogeneous data in a broad range of applications. The restricted maximum likelihood method is often preferred to estimate co-variance parameters in such models due to its unbiased estimation of the underlying variance parameters. The restricted log-likelihood function involves log determinants of a complicated co-variance matrix. An efficient statistical estimate of the underlying model parameters and quantifying the accuracy of the estimation requires the first derivatives and the second derivatives of the restricted log-likelihood function, i.e., the observed information. Standard approaches to compute the observed information and its expectation, the Fisher information, is computationally prohibitive for linear mixed models with thousands random and fixed effects. Customized algorithms are of highly demand to keep mixed models analysis scalable for increasing high-throughput heterogeneous data sets. In this paper, we will explain why the multi-frontal solver is preferred for such a class of problems, we explore how to leverage an averaged information splitting technique and dedicate matrix transform to significantly reduce computations, and how to use the multi-frontal solvers on linear systems with multiple right-hand sides to accelerate computing. This research is stimulated by a knowledge transfer project by EPSRC, VSN International Ltd and the University of Oxford, was supervised by Andy Wathen.