Sparse Days 2018

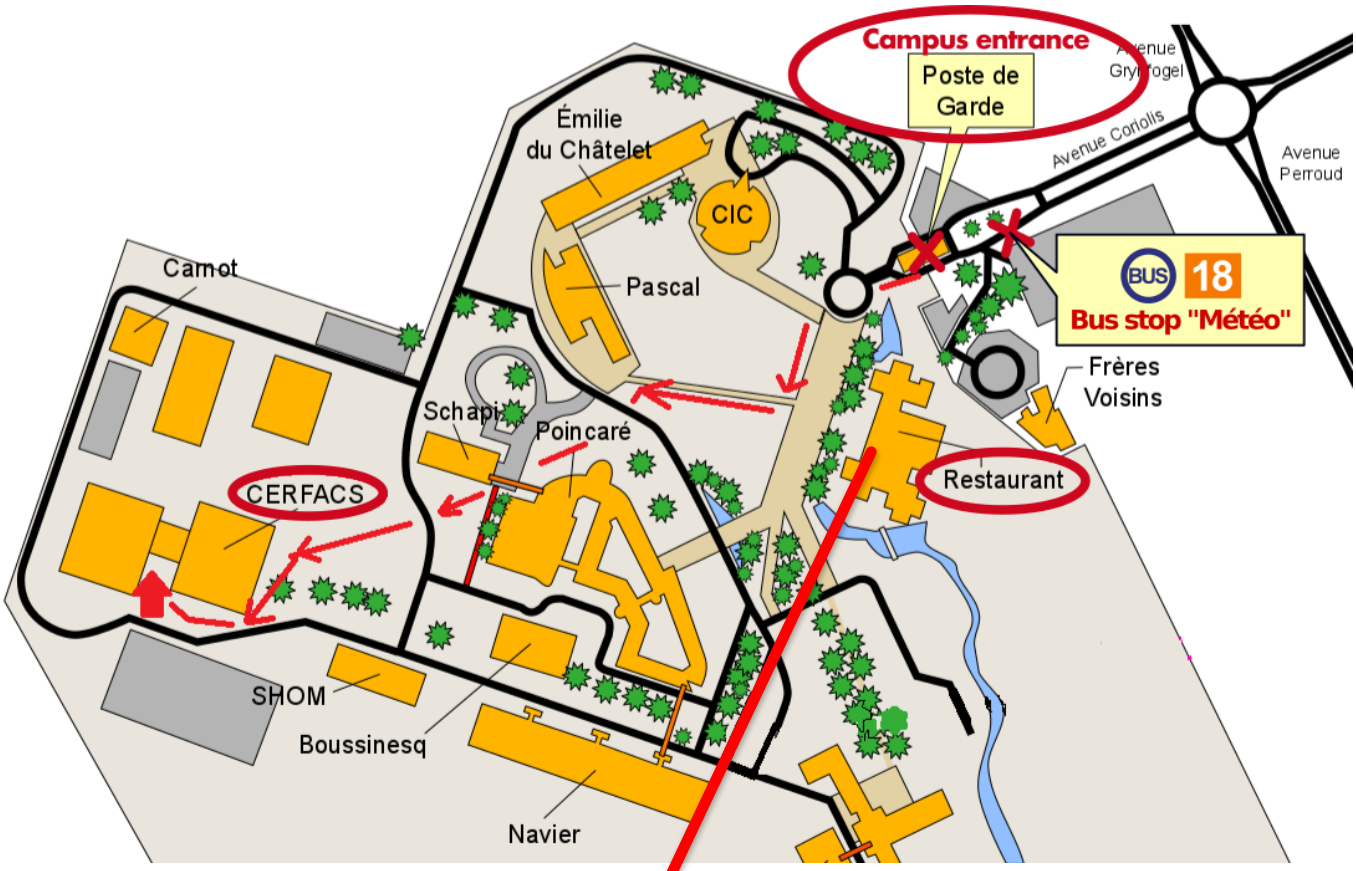
September 27th – 28th, 2018

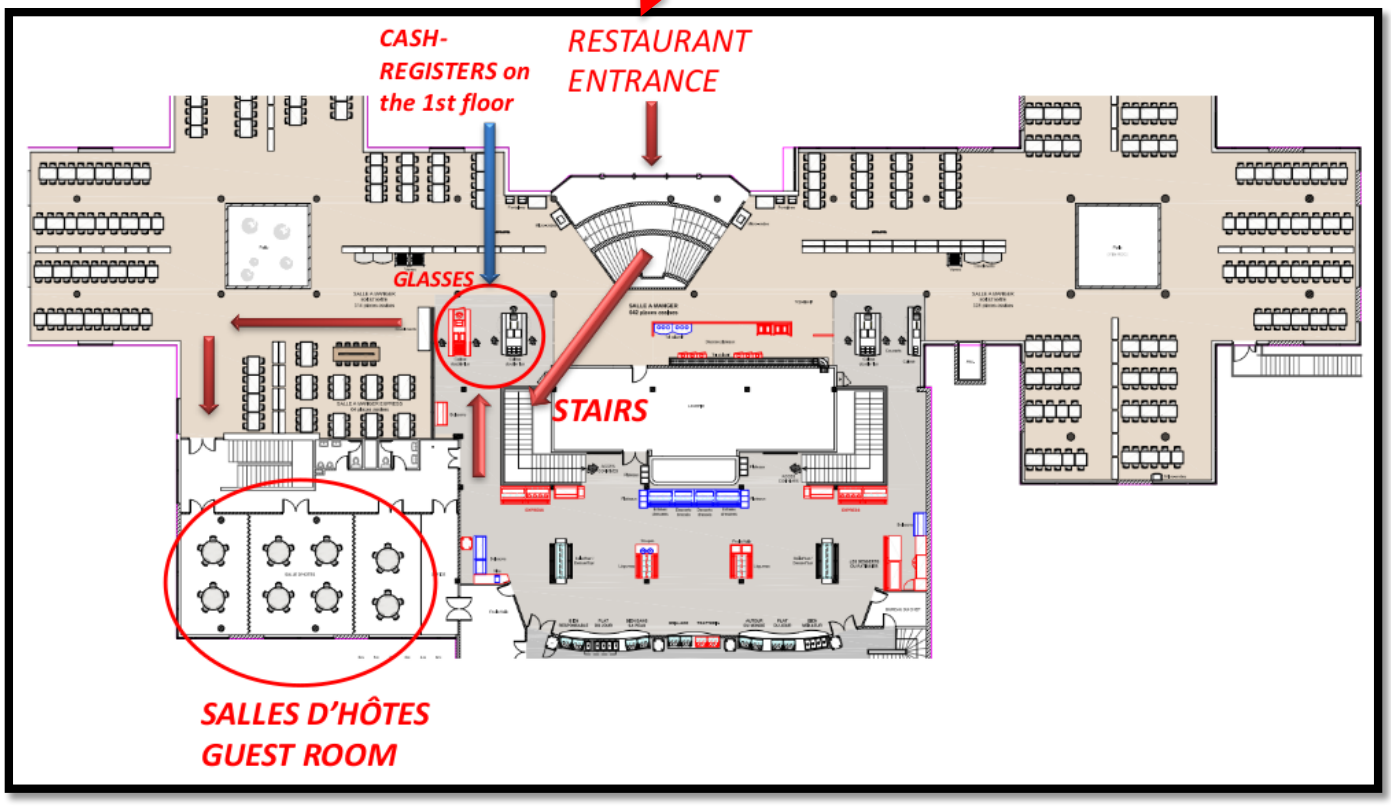
CERFACS, Toulouse, France

CERFACS

CENTRE EUROPÉEN DE RECHERCHE ET DE FORMATION AVANCÉE EN CALCUL SCIENTIFIQUE

Welcome to Cerfacs





Sparse Days programme
Thursday, September 27th 2018

09:00 – 10:15 **Registration and coffee**

10:15 – 10:30 **Welcome message**

10:30 – 12:10 **Session 1 – chair: Iain Duff**

10:30 – 10:55 What is happening in non-negative matrix factorization?
Man Shun ANG (Université de Mons, Belgium)

10:55 – 11:20 A New Preconditioner for Low-Precision Block Low-Rank Multifrontal Solvers
Theo MARY (University of Manchester, UK)

11:20 – 11:45 Quantum circuits synthesis using Householder transformations
Timothée GOUBAULT DE BRUGIERE (LRI - Université Paris-Sud, France)

11:45 – 12:10 Hierarchical Symbolic Factorization for Sparse Matrices
Aurélien FALCO (Inria Bordeaux, France)

12:10 – 14:00 **Lunch break**

14:00 – 15:40 **Session 2 – chair: Marc Baboulin**

14:00 – 14:25 A distributed-memory parallel approximation of maximum weight perfect bipartite matching
Johannes LANGGUTH (Simula research laboratory, Norway)

14:25 – 14:50 Task-based sparse direct solvers for symmetric systems in the NLAFET Project
Florent LOPEZ (Rutherford Appleton Laboratory, UK)

14:50 – 15:15 Parallel Implementations for Solving Tridiagonal Systems
Pedro VALERO-LARA (Barcelona Supercomputing Center, Spain)

15:15 – 15:40 OpenSPARSE: An Open Platform for High Performance Sparse Basic Linear Algebra Subprograms
Weifeng LIU (Norwegian University of Science and Technology, Norway)

15:40 – 16:00 **Coffee break**

16:00 – 17:15 **Session 3 – chair: Daniel Ruiz**

16:00 – 16:25 Dissection solver for higher precision arithmetic by inner iterative refinement
Atsushi SUZUKI (Cybermedia Center - Osaka University, Japan)

16:25 – 16:50 On an efficient parallel implementation of adaptive FETI-DP with load balancing
Martin KÜHN (University of Cologne, Germany)

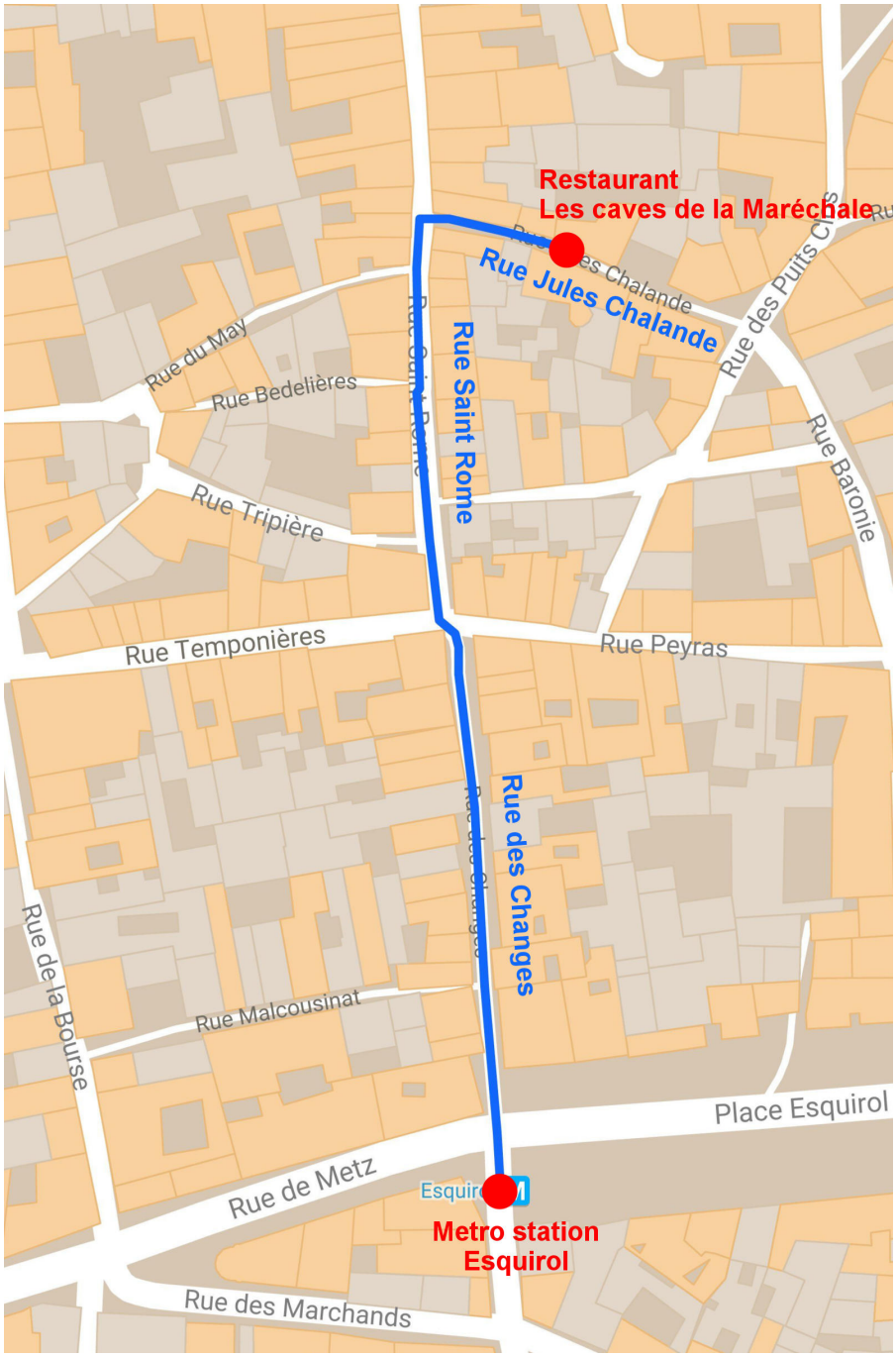
16:50 – 17:15 High-Order Methods for Parabolic Equations in Multiple Space Dimensions for Option Pricing Problems
Matthias EHRHARDT (University of Wuppertal, Germany)

19 :45 **Reception dinner**
Les Caves de la Maréchale, 3 rue Jules Chalande, Toulouse

Reception dinner at Les Caves de la Maréchale
3 rue Jules Chalande, Toulouse
Thursday, September 27th 2018, 19:45

Directions from Météo-France to the restaurant:

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.



Friday, September 28th 2018

09:30 – 10:45 Session 4 – chair: Selime Gürol

09:30 – 09:55 Numerical Stability of s-step Enlarged Krylov Subspace Conjugate Gradient methods
Sophie MOUFAWAD (American University of Beirut, Lebanon)

09:55 – 10:20 Efficient algebraic coarse spaces
Hussam AL DAAS (Inria Paris, France)

10:20 – 10:45 Parallel preconditioning for time-dependent PDE problems
Andy WATHEN (Oxford University, UK)

10:45 – 11:05 Coffee break

11:05 – 12:20 Session 5 – chair: Uli Rüde

11:05 – 11:30 AMG preconditioning with Parallel Aggregation based on Compatible Weighted Matching
Salvatore FILIPPONE (Cranfield University, UK)

11:30 – 11:55 aSP-AMG: adaptive Smoothing and Prolongation Algebraic MultiGrid
Victor MAGRI (University of Padua, Italy)

11:55 – 12:20 Fast iterative solvers for robust discretisations
Frank HÜLSEMANN (EDF R&D, France)

12:20 – 14:00 Lunch break

14:00 – 15:15 Session 6 – chair: Luc Giraud

14:00 – 14:25 Performance portable parallel CP-APR tensor decompositions
Keita TERANISHI (Sandia National Laboratories, USA)

14:25 – 14:50 Communication-Avoiding Sparse-Matrix Primitives for Parallel Machine Learning
Aydin BULUC (Lawrence Berkeley National Laboratory, USA)

14:50 – 15:15 Impacts of Three Soft-Fault Models on Hybrid Parallel Asynchronous Jacobi
Masha SOSONKINA (Old Dominion University, USA)

15:15 – 15:35 Coffee break

15:35 – 16:50 Session 7 – chair: Iain Duff

15:35 – 16:00 Exploiting ultra-sparsity in the revised simplex method
Julian HALL (University of Edinburgh, UK)

16:00 – 16:25 Preconditioned linear solvers in CMB data analysis
Jan PAPEZ (Inria Paris, France)

16:25 – 16:50 A Tale of Woe
Robert LUCAS (LSTC, USA)

Abstracts

Efficient algebraic coarse spaces

Hussam AL DAAS (Inria Paris, France)

We present a class of robust and fully algebraic two-level preconditioners for SPD matrices. A notion of algebraic local SPSD splitting of an SPD matrix is introduced and we give a characterization of this splitting. This splitting leads to construct *algebraically and locally* a class of efficient coarse spaces which bound the spectral condition number of the preconditioned system by a number defined a priori. We also introduce the τ-filtering subspace. This concept helps compare the dimension minimality of coarse spaces. Some PDEs-dependant preconditioners correspond to a special case. The examples of the theoretical algebraic coarse spaces are not practical due to the expensive construction of the corresponding splitting. We propose a heuristic approximation that is not costly. Numerical experiments illustrate the efficiency of the proposed method.

What is happening in non-negative matrix factorization?

Man Shun ANG (Université de Mons, Belgium)

Given a non-negative input matrix X, the goal of the Non-negative matrix factorization (NMF) is to decompose X into two (smaller) matrices W and H such that the product WH fits X under some metric distances such as the Frobenius norm and the Beta divergence. The non-negativity in NMF makes NMF finds many applications, due to the fact that the decomposed factors of NMF enjoy a more interpretability than the factors obtained by other methods. Despite the success of NMF in many applications, the NMF model itself is an ill-posed, underdetermined problem. Because of this, different new formulations of NMF are proposed in the past decade.

This talk discuss what is going on NMF in this direction: from the classical separability condition, to the relaxed sufficiently scattered condition, minimum volume and the more.

Communication-Avoiding Sparse-Matrix Primitives for Parallel Machine Learning

Aydin BULUC (Lawrence Berkeley National Laboratory, USA)

Machine learning (ML) has proven to be a powerful tool for analyzing business, social and scientific data sets. Data from scientific domains where ML methods excel are often sparse with many missing entries. Additionally, many ML methods rely on numerical optimization algorithms, which themselves are based on sparse matrix operations. Consequently, scalable sparse linear algebra software is sorely needed for ML due to sparse datasets or the need to enforce output (and model) sparsity for avoiding overfitting and increasing interpretability.

The challenges of ML for science problems include extreme-scale data volumes and data rates, necessitating parallel ML algorithms that will run on exascale architectures. Due to sparsity, popular implementations of common ML algorithms struggle to efficiently harness the capabilities of large-scale parallel computers. One prevalent problem is the increasingly dominant cost of communication. In this talk, I will describe our recent work on distributed-memory parallelization of prevalent ML problems such as graphical model estimation and flow-based clustering. I will highlight the importance of communication-avoiding sparse matrix primitives to achieve scalability in these problems. I will conclude with open problems and future directions.

High-Order Methods for Parabolic Equations in Multiple Space Dimensions for Option Pricing Problems

Matthias EHRHARDT (University of Wuppertal, Germany)

In this talk we discuss higher-order spatial discretization methods using finite differences and pseudo-spectral methods and demonstrate how to use it in a sparse grid combination technique. Next, for the time discretization we propose alternating direction implicit (ADI) schemes and analyse its stability. We conclude with two applications to financial engineering partial differential equations: Basket-Options in the Black-Scholes model and European Plain-Vanilla options under Stochastic Volatility.

Hierarchical Symbolic Factorization for Sparse Matrices

Aurélien FALCO (Inria Bordeaux, France)

Hierarchical methods based on low-rank compression have drastically reduced computational requirements for the solution of dense linear systems over the last two decades. For sparse linear systems, more common in numerical simulation, their application remains a challenge which has been studied both by the community of hierarchical matrices and the community of sparse matrices. On one hand, the first step taken by the community of hierarchical matrices most often takes advantage of the sparsity of the problem through the use of nested dissection. While this method benefits from the resulting hierarchical structure, it is not, however, as efficient as sparse solvers regarding the exploitation of zeros and the structural separation of zeros from non-zeros. On the other hand, sparse linear systems can be decomposed as a sequence of smaller dense operations, enticing sparse solvers to use this property to borrow compression techniques from hierarchical methods to reduce the computational cost of these elementary operations. Nonetheless, the globally hierarchical structure may be lost if the compression of hierarchical methods is used only locally on dense submatrices. In this presentation, we will review the main techniques that have been employed by both those communities, trying to highlight their shared properties and their respective limits with a special emphasis on studies that have aimed at bridging the gap between them. This review motivates the introduction of a class of hierarchical algorithms performing a symbolic factorization at different levels of the cluster (or block elimination) tree and questions whether supernodes shall have (hierarchically) consistent data structures. Experiments with a test application (representative of Airbus' targeted industrial application in aeroacoustics) will illustrate our discussion.

AMG preconditioning with Parallel Aggregation based on Compatible Weighted Matching

Salvatore FILIPPONE (Cranfield University, UK)

Recently we introduced the BootCMatch software package for the construction of AMG hierarchies based on weighted graph matching [1]. In this talk we will discuss some early experience with integrating this aggregation approach with the parallel AMG preconditioners of the MLD2P4 package. We compare the effectiveness and computational cost of the various matching strategies available in the context of linear systems of medium to large size arising from applications in the H2020 EoCoE project, and discuss options for future development and improvements.

[1] P. D'Ambra, S. Filippone, P. Vassilievski. BootCMatch: a software package for bootstrap AMG based on graph weighted matching, ACM TOMS, Vol 44, 30:1-39:25, 2018.

Quantum circuits synthesis using Householder transformations

Timothée GOUBAULT DE BRUGIERE (LRI - Université Paris-Sud, France)

In quantum computing, operators can be expressed via unitary matrices that can be formalized through so-called quantum circuits composed of basic operations - quantum gates. Then a major issue in quantum compiling is to decompose a large unitary matrix into a series of elementary matrices. Each elementary matrix needs to be easily implementable into a sequence of quantum gates in order to get a viable quantum circuit. During this compilation process we need to optimize both quantum and classical resources that correspond to the number of quantum gates in the resulting circuit (requiring quantum resources) and the computational time to obtain this circuit (classical resources). The first criterion has been optimized in the past but the second much less. In this presentation, we adapt the QR factorization of unitary matrices to express it as a product of Householder transformations and a diagonal matrix, for which simple quantum implementations are known. We express the computational costs (in flop and gate counts) and study the performance by comparing them to existing techniques from the literature. We highlight an inevitable trade-off between the time spent in the classical and quantum hardware, respectively.

Exploiting ultra-sparsity in the revised simplex method

Julian HALL (University of Edinburgh, UK)

The revised simplex method is generally the technique of choice when solving large sparse linear programming problems, and computational efficiency is driven by the need to compute rows and columns of the standard simplex tableau. Problems for which these vectors are sparse are said to exhibit hyper-sparsity, and techniques to exploit this computationally were developed about 20 years ago. However, for some problems the degree of hyper-sparsity is such that these techniques may be cache inefficient. This talk will discuss techniques by which such ultra-sparsity may be exploited with the aim of gaining further performance improvement.

Fast iterative solvers for robust discretisations

Frank HÜLSEMANN (EDF R&D, France)

Recently developed robust discretisation methods such as Compatible Discrete Operators (CDO) and the Hybrid High Order (HHO) framework give rise to linear systems that pose a number of challenges for existing (algebraic) multigrid methods. While mesh quality matters, some convergence issues appear already for the diffusion operator on uniform meshes. This presentation summarizes available convergence results and the work in progress to analyse the reasons for the – so far – unsatisfactory solver behaviour.

On an efficient parallel implementation of adaptive FETI-DP with load balancing

Martin KÜHN (University of Cologne, Germany)

Domain decomposition methods such as FETI-DP (Finite Element Tearing and Interconnecting - Dual Primal) and BDDC (Balancing Domain Decomposition by Constraints) are highly scalable parallel solvers for large sparse systems obtained from the discretization of partial differential equations (PDEs).

However, the convergence behavior of FETI-DP and BDDC methods with a standard coarse space highly depends on the parameters of the underlying PDE. The convergence rate of both methods can deteriorate significantly if composite materials are considered. In such cases, problem-dependent (or adaptive) coarse spaces offer a remedy. In adaptive methods, difficulties arisen from highly heterogeneous materials are detected automatically by solving local generalized eigenvalue problems and an adaptive coarse space is set up. These methods are thus characterized by great robustness.

Though, for an efficient parallel implementation, different issues such as load imbalances and the solution of unnecessary eigenvalue problems have to be avoided to reduce the computational overhead in the set up phase. We will present details of the set up of the adaptive method to implement the coarse space enrichment efficiently in a parallel context.

We will present weak and strong scaling results to show the good parallel scalability of our method.

A distributed-memory parallel approximation of maximum weight perfect bipartite matching

Johannes LANGGUTH (Simula research laboratory, Norway)

We discuss efficient parallel approximation algorithm for the problem of maximum weight perfect matching in bipartite graphs, i.e. the problem of finding a set of non-adjacent edges that covers all vertices and has maximum weight. This problem differs from the maximum weight matching problem, for which scalable approximation algorithms are known. It is primarily motivated by finding good pivots in scalable sparse direct solvers before factorization where sequential implementations of maximum weight perfect matching algorithms, are generally used due to the lack of scalable alternatives. To overcome this limitation, we propose a parallel distributed memory algorithm and discuss its approximation properties.

OpenSPARSE: An Open Platform for High Performance Sparse Basic Linear Algebra Subprograms

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

In the past decades, much research has been focusing on designing new data structures and efficient algorithms for sparse basic linear algebra subprograms (BLAS), such as sparse matrix-vector multiplication, sparse triangular solve and sparse matrix-matrix multiplication. However, only very limited research results have been incorporated in widely-used sparse libraries, and have benefited real-world applications. This talk will present our recent ideas and effort to establish an open platform that can bridge the gap between high performance sparse BLAS research and mathematical software development.

The solution of sparse linear systems in the NLAFET Project

Florent LOPEZ (Rutherford Appleton Laboratory, UK)

We discuss work done in Task 3.2 within the EU Project NLAFET.

Many applications in science and engineering require the solution of large sparse linear systems of equations. For solving such problems, direct methods are frequently employed because of their robustness, accuracy and usability as black-box solvers. As modern architectures become more and more complex, with an increasing number of cores per chip, a deeper memory hierarchy and the integration of accelerators such as GPUs, it becomes more and more challenging to exploit the computational capabilities of such machines for sparse matrix factorization algorithms.

We first focus on the symmetric positive-definite case where we describe the parallelization of the solve phase of our SpLLT solver using OpenMP tasks. We show experimentally that it performs better that the state-of-the-art solvers PARDISO and PaStiX. Although the cost of the solve phase often seems marginal compared to the factorization, we show that it can give substantial benefits when the solve is performed many times within an iterative scheme. In particular, we show a dramatic performance improvement for the Enlarged Conjugate Gradient (ECG) solver of Inria when using SpLLT instead of PARDISO. The ECG solver has been developed by Inria as part of Workpackage 4 in the NLAFET Project.

We then consider the symmetric indefinite case for which we have developed the SpLDLT solver. This solver implements a DAG-based multifrontal method that relies on a APTP (A Posteriori Threshold Pivoting) strategy and uses the StarPU runtime system for implementing the parallel version. We show that our solver compares favourably to the state-of-the-art solvers HSL_MA86, HSL_MA97 and PARDISO in a multicore environment and we discuss the benefits of our approach for exploiting heterogeneity in the the context of GPU-accelerated systems.

This is joint work with Sebastien Cayrols, Iain Duff, and Stojce Nakov.

A Tale of Woe

Robert LUCAS (LSTC, USA)

Only time and resource constraints limit the size and complexity of the implicit analyses that LS-DYNA user's would like to perform. Cray, LSTC, NCSA, and Rolls Royce formed a partnership to explore the future of implicit computations as both the finite element models and the systems they run on increase in scale. To facilitate this work, Rolls Royce created a family of dummy engine models, using solid elements, with as many as 200,000,000 degrees of freedom. These are the largest implicit LS_DYNA models we are aware of. NCSA ran these with specialized LS-DYNA variants, generated by Cray, using their Blue Waters machine, a hybrid Cray XE/XK system with 360,000 AMD cores. CrayPat and novel NCSA-developed runtime monitoring tools were used to identify both processing and memory bottlenecks that revealed themselves when the number of processors in increased by an order-of-magnitude beyond that familiar to today's developers and users. This talk will discuss the challenges encountered, enhancements made to LS-DYNA, and the results when extending the limits both in terms of the scale of the model, and the number of processors. This is ongoing work, and we will conclude by discussing the path forward that has been illuminated.

aSP-AMG: adaptive Smoothing and Prolongation Algebraic MultiGrid

Victor MAGRI (University of Padua, Italy)

The numerical simulation of modern engineering problems can easily incorporate millions or even billions of degrees of freedom. In several applications, these simulations require the solution of sparse linear systems of equations, and algebraic multigrid (AMG) methods are often standard choices as iterative solvers or preconditioners. Despite carrying the name "algebraic", most of these methods still rely on additional information other than the global assembled sparse matrix, for instance, the knowledge of the operator near kernel. This fact somewhat limits their applicability as black-box solvers. In this presentation, we introduce a novel adaptive AMG approach featuring the adaptive Factored Sparse Approximate Inverse (aFSAI) method as a flexible smoother; two strategies for uncovering the near-null space of the system matrix and, lastly, a new approach to compute the prolongation operator dynamically. We assess the performance of the proposed AMG through the solution of a set of model problems along with real-world engineering applications. Moreover, comparisons are made with the aFSAI and BoomerAMG preconditioners, showing that our new method proves to be superior to the first method and with performance comparable to the second one while being especially attractive (see Figure 1) for the solution of linear elasticity problems.

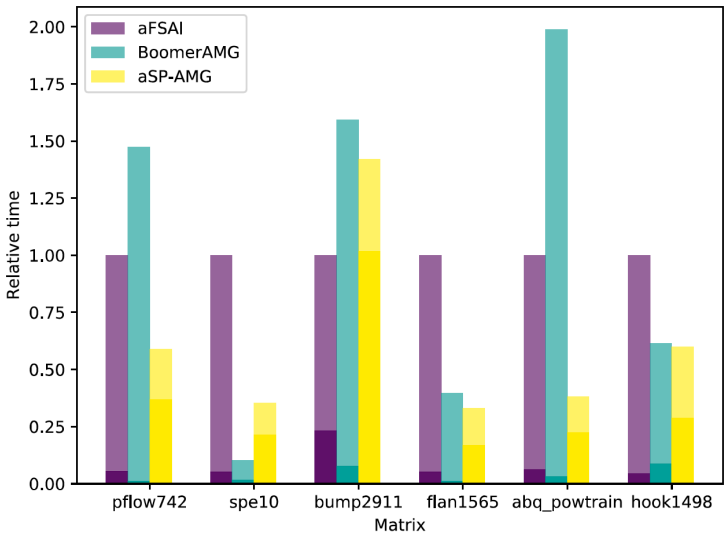


Figure 1: relative solution time comparison among aFSAI, BoomerAMG and aSP-AMG when used as preconditioners to CG in the solution of real-world engineering problems. The dark sections denote the preconditioner setup time while the light ones stand for the solution time, both quantities are normalized with the total (setup + solution) time showed by aFSAI. The first two test cases refer to the simulation of fluid flow in heterogeneous oil reservoirs while the remaining ones are elastostatics problems.

A New Preconditioner for Low-Precision Block Low-Rank Multifrontal Solvers

Theo MARY (University of Manchester, UK)

We consider the solution of sparse linear systems Ax=b via the computation of a low precision LU factorization A ≈ LU used to precondition some iterative method. We have observed that the error matrix $E=U^{-1}L^{-1}A - I$ is often numerically low-rank when A is ill-conditioned. We propose a novel preconditioner based on exploiting this low-rank property of the error to accelerate the convergence of iterative methods. In our numerical experiments, we showcase the performance of this new preconditioner to solve a variety of real-life problems with several types of approximate LU factorizations, with a special focus on low-precision block low-rank (BLR) multifrontal solvers.

Numerical Stability of s-step Enlarged Krylov Subspace Conjugate Gradient methods

Sophie MOUFAWAD (American University of Beirut, Lebanon)

Recently, enlarged Krylov subspace methods, that consists of enlarging the Krylov subspace by a maximum of t vectors per iteration based on the domain decomposition of the graph of A, were introduced in the aim of reducing communication when solving systems of linear equations Ax=b. In this talk, the s-step enlarged Krylov subspace Conjugate Gradient methods are introduced, whereby s iterations of the enlarged Conjugate Gradient methods are merged in one iteration. The numerical stability of these s-step methods is studied, and several numerically stable versions are proposed. Similarly to the enlarged Krylov subspace methods, the s-step enlarged Krylov subspace methods have a faster convergence than Krylov methods, in terms of iterations. Moreover, by computing st basis vectors of the enlarged Krylov subspace $K_{k,t}(A,r_0)$ at the beginning of each s-step iteration, communication is further reduced. Thus, the introduced methods are parallelizable with less communication, with respect to their corresponding enlarged versions and to Conjugate Gradient.

Preconditioned linear solvers in CMB data analysis

Jan PAPEZ (Inria Paris, France)

Studies of the Cosmic Microwave Background (CMB) anisotropies have been driving the progress in our understanding of the Universe for more than 20 years. The current and forthcoming CMB observatories are expected to deliver unprecedented insights about the Universe's beginning and evolution, producing enormous data sets of size $O(10^{15})$ and thus calling for advanced, high performance data analysis techniques and efficient algebraic solvers.
We present several applications in CMB data analysis that result in linear algebraic problems or sequences of problems with matrices that, properly represented, are large but very sparse. Then we discuss the state-of-the-art algebraic solvers as well as an alternative technique, which has become recently popular in the CMB field.
This is a joint work with Laura Grigori (INRIA Paris) and Radek Stompor (Université Paris 7 Denis Diderot)

Impacts of Three Soft-Fault Models on Hybrid Parallel Asynchronous Jacobi

Masha SOSONKINA (Old Dominion University, USA)

This study seeks to understand the soft error vulnerability of asynchronous iterative methods, with a focus on stationary iterative Jacobi solver. The implementations are based on hybrid parallelism where the computational work is distributed over multiple nodes using MPI and parallelized on each node using OpenMP. A series of experiments is conducted to measure the impact of an undetected soft fault on Jacobi and to compare and contrast several techniques for simulating the occurrence of a fault and then recovering from the effects of the faults. The data shows that the two numerical soft-fault models tested here more consistently than a “bit-flip” model produce bad enough behavior to test a variety of recovery strategies, such as those based on partial checkpointing.

Dissection solver for higher precision arithmetic by inner iterative refinement

Atsushi SUZUKI (Cybermedia Center - Osaka University, Japan)

A parallel sparse direct solver code, "Dissection" performs recursive generation and factorization of Schur complement matrices following nested-dissection ordering with postponing factorization strategy by which suspicious null pivots are excluded from the elimination tree. For some matrices with high condition number whose data is given in double precision, factorization process and forward/backward substitution process in quadruple precision are necessary to obtain solution in appropriate accuracy. "Dissection" is written by C++ using template facility and then as a result, it is possible to perform the hole procedure in quadruple precision using "double-double" arithmetic. However, due to lack of optimized BLAS 3 library in quadruple precision, computational time by "double-double" is about 100 times slower than by standard "double". To reduce huge cost of quadruple precision arithmetic, an iterative refinement technique is applied to generate the last Schur complement matrix consisting of postponed entries and some invertible entries from the top of the elimination tree and LDU factorization in quadruple precision is applied to that matrix having high condition number. The iterative refinement procedure consists of multiplication of sparse matrix with excluded some entries to multiple right-hand sides in quadruple precision and solution of multiple-RHS residual in double precision. This hybrid method works better than global iterative refinement process with double-precision factorization of the whole matrix. Numerical results on matrices from a semi-conductor problem will be shown in comparison with full quadruple precision arithmetic.
This work is in collaboration with François-Xavier Roux (ONERA/LJLL).

Performance portable parallel CP-APR tensor decompositions

Keita TERANISHI (Sandia National Laboratories, USA)

Tensors have found utility in a wide range of applications, such as chemometrics, network traffic analysis, neuroscience, and signal processing. Many of these data science applications have increasingly large amounts of data to process and require high-performance methods to provide a reasonable turnaround time for analysts. Sparse tensor decomposition is a tool that allows analysts to explore a compact representation (low-rank models) of high-dimensional data sets, expose patterns that may not be apparent in the raw data, and extract useful information from the large amount of initial data. In this work, we consider decomposition of sparse count data using CANDECOMP-PARAFAC Alternating Poisson Regression (CP-APR).
Unlike the Alternating Least Square (ALS) version, CP-APR algorithm involves non-trivial constraint optimization of nonlinear and nonconvex function, which contributes to the slow adaptation to high performance computing (HPC) systems. The recent studies by Kolda et al. suggest multiple variants of CP-APR algorithms amenable to data and task parallelism together, but their parallel implementation involves several challenges due to the continuing trend toward a wide variety HPC system architecture and its programming models.
To this end, we have implemented a production-quality sparse tensor decomposition code, named SparTen, in C++ using Kokkos as a hardware abstraction layer. By using Kokkos, we have been able to develop a single code base and achieve good performance on each architecture. Additionally, SparTen is templated on several data types that allow for the use of mixed precision to allow the user to tune performance and accuracy for specific applications. In this presentation, we will use SparTen as a case study to document the performance gains, performance/accuracy tradeoffs of mixed precision in this application, development effort, and discuss the level of performance portability achieved. Performance profiling results from each of these architectures will be shared to highlight difficulties of efficiently processing sparse, unstructured data. By combining these results with an analysis of each hardware architecture, we will discuss some insights for improved use of the available cache hierarchy, potential costs/benefits of analyzing the underlying sparsity pattern of the input data as a preprocessing step, critical aspects of these hardware architectures that allow for improved performance in sparse tensor applications, and where remaining performance may still have been left on the table due to having single algorithm implementations on diverging hardware architectures.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Parallel Implementations for Solving Tridiagonal Systems

Pedro VALERO-LARA (Barcelona Supercomputing Center, Spain)

Many problems of industrial and scientific interest require the solving of tridiagonal linear systems. This work presents several implementations for the parallel solving of tridiagonal systems on multi-core and GPU architectures, using OmpSs and CUDA respectively. Depending of the applications, the size and the number of tridiagonal systems to be computed is completely different, due to this, we present two different approaches, one for solving a batch of independent tridiagonal systems, and one for solving a large single tridiagonal system. While on NVIDIA GPUs, we focus on solving a batch of tridiagonal systems, on multi-core processors, we focus on solving one single large tridiagonal system. Although different strategies and implementations are explored and presented, depending on the target problem and platform, the strategy used for the parallelization is based on the use of the two most popular existing algorithms, PCR and Thomas. The Thomas algorithm, which cannot be parallelized, is the optimal algorithm in terms of number of floating point operations. The PCR algorithm is the most popular parallel method, but it is more computationally expensive than Thomas. For the parallelization of large tridiagonal systems on multi-core, the method proposed consists of applying the PCR algorithm to break down one large tridiagonal system into a set of smaller and independent ones. In a second step, these independent systems are concurrently solved using Thomas. We also present an analytical study of which is the best point to switch from PCR to Thomas. The performance evaluation shows that the best implementation achieves a peak speedup of 4 with respect to the Intel MKL counterpart routine. We propose a new implementation (cuThomasBatch) based on the Thomas algorithm for the solving of a batch of independent tridiagonal systems. As commented before, the Thomas algorithm is sequential, and so a coarse-grained approach is implemented where one CUDA thread solves a complete tridiagonal system instead of one CUDA block as in the routine of the cuSparse library, gtsvStridedBatch. To achieve a good scalability using this approach, it is necessary to carry out a transformation in the way that the inputs are stored in memory to exploit coalescence (contiguous threads access to contiguous memory locations). The results given in this study prove that the implementations carried out in this work are able to beat the reference code gtsvStridedBatch, being up to 5× (in double precision) and 6× (in single precision) faster using the NVIDIA GPU Pascal P100 architecture.

Parallel preconditioning for time-dependent PDE problems

Andy WATHEN (Oxford University, UK)

Monolithic (or all-at-once) discretizations of evolutionary problems most often give rise to nonsymmetric linear(ised) systems of equations which can be of very large dimension for PDE problems. In this talk we will describe preconditioners for such systems with guaranteed fast convergence via use of MINRES (not LSQR or CGNE) or GMRES. These results apply with standard time-stepping schemes and relate to Toeplitz matrix technology and preconditioning via circulants using the FFT. Simple parallel computational results will be shown for the heat equation and the wave equation.

This is joint work with Elle McDonald (CSIRO, Australia), Jennifer Pestana (Strathclyde University, UK) and Anthony Goddard (Oxford University, UK)