Centre International
de Mathématiques et d'Informatique
TOULOUSE

BOOK OF ABSTRACTS

HPC Semester

Sparse Days in Saint - Girons III

June 29th – July 2nd, 2015
Saint - Girons, France

About the CIMI semester

High Performance Linear and Nonlinear Methods for Large Scale Applications

The aim of this CIMI semester is to take advantage of the opportunity of an international conference in the field of high performance computing and linear algebra, that is organized about every ten years in St Girons (previous dates were July 1994 and June 2003), by organizing a series of satellite events, workshops, and visits that will benefit both applied mathematics and computer science research activities.

Solving sparse linear systems is often the most time-consuming computation in many large-scale computer simulations in science and engineering, such as computational fluid dynamics, structural analysis, and design of new materials in nanoscience. In recent years this has extended further to such diverse areas as life sciences and finance. If the solution of linear systems of order of a few million was a challenging task ten years ago, a linear system can today have more than a billion variables for the problems arising in many of the previously mentioned applications. At the same time, the power and efficiency of computing systems have been growing very rapidly largely because of the introduction of new processor technologies such as multicore and accelerators such as GPUs. As a result, modern computers are characterized by complex architectures with deep memory hierarchies, extremely high degrees of parallelism and a heterogeneous mixture of computing devices. Research in sparse linear algebra is thus naturally at the intersection between many fields of research such as numerical linear algebra, combinatorial science for large graph processing, and parallel computing. This diversity has been the main motivation for organizing, around the abovementioned international event, four satellite activities that will involve numerical issues, parallelism and combinatorial science.

Events of the semester:

- **Sparse Days in St Girons III**, June 29th - July 2nd 2015 is an international conference and will be the central event of the semester.
  Conference general chairs: James Demmel (University of Berkeley), Jack Dongarra (University of Tennessee), Iain Duff (RAL and CERFACS-IRIT joint lab), and Patrick Amestoy (INPT-IRIT). The local organizer chair is Pierre-Henri Cros (IRIT, Université de Toulouse).

- In Toulouse at INP-ENSEEIHT (4:00pm-6:00pm), a large audience conference by **Jack Dongarra** (CIMI advisory board member) and **Satoshi Matsuoka** on July 2nd 2015:
  Current trends for high performance computing and challenges for the future

- **Workshops:**

  1. **Low-rank approximations for high performance sparse solvers**
     Toulouse, June 24th - June 26th 2015
     Local organizers: Alfredo Buttari (CNRS-IRIT) and David Titley-Peloquin (CERFACS-IRIT joint lab)
  2. **Innovative clustering methods for large graphs and block methods**
     Toulouse, July 6th - July 8th 2015
     Local organizers: Sandrine Mouysset (UPS-IRIT) and Daniel Ruiz (INPT-IRIT).
  3. **Parallel methods for time dependent problems**
     Toulouse, January 11th-12th 2016
     Local organizers: Xavier Vasseur (CERFACS-IRIT joint lab) and Serge Gratton (CERFACS-IRIT joint lab)
  4. **Advances in optimization with application to data assimilation**
     Toulouse, January 13th-15th 2016
     Local organizers: Serge Gratton (CERFACS-IRIT joint lab) and Ehouarn Simon (INPT-IRIT)

Sparse Days in St Girons III

In July 1994, a meeting was held in St Girons, Ariège, to launch the CERFACS International Linear Algebra Year. It was such a success that a second Sparse Days at St Girons was held in June 2003 and it was felt that it was appropriate to schedule the central event of the CIMI semester, Sparse Days in St Girons III, in the same location. The conference is organized by IRIT and CERFACS researchers with an international management committee comprising general chairs

- James Demmel (University of Berkeley),
- Jack Dongarra (University of Tennessee),
- Iain Duff (RAL and CERFACS-IRIT common lab), and
- Patrick Amestoy (INPT-IRIT, Université de Toulouse)

The local organizer chair is Pierre-Henri Cros (IRIT, Université de Toulouse).

In any large-scale numerical computation in the context of large, complex and hierarchical computers, one has to understand the algorithmic choices that have to be made. In this context, both sparse linear algebra kernels together with dense linear kernels are the key to enable other numerical libraries in the software stack to achieve high computational efficiency. The focus of the conference in St Girons will be on these kernel topics although there will be dedicated sessions on optimization and on the themes of the four workshops of the CIMI semester:

- Low-rank approximations for high performance sparse solvers
- Innovative clustering methods for large graphs and block methods
- Parallel methods for time dependent problems
- Advances in optimization with application to data assimilation

The St Girons conference will be followed on July 2nd 2015 at INP-ENSEEIHT (4:00pm-6:00pm) by the large audience conference by **Jack Dongarra** and **Satoshi Matsuoka** *"Current trends for high performance computing and challenges for the future"*. During the conference the speakers will examine how high performance computing has changed over the last ten years and will look toward the future in terms of trends towards, and beyond, so-called extreme "exascale" computing as well as "big data", that pose unprecedented challenges both in terms of hardware and software. A new generation of software libraries and algorithms are needed for the effective and reliable use of computing and data resources in complex, dynamic, and (wide area) distributed and parallel environments.

About the town of St Girons

The little town of Saint Girons, nestled close to the Pyrénées stands out for its exceptional environmental quality. Our Couserans area is the crossroads and economic centre of the Regional Natural Park swarming with cultural and natural resources.

Visitors can practice various sporting activities: hiking, cycling, climbing, canyoning, gliding, skiing, speleology. They will enjoy our traditional food and cooking with natural local products. The historical monuments and the Papal city of Saint-Lizier are worth visiting and attract a lot of tourists.

But Science and Research are not outdone: in Moulis, the CNRS (Scientific Research National Centre) is the best place in the world for biodiversity. L'*Etang de l'Hers* is a unique geological spot with the Lherzolite. Alexandre Grothendieck, a very famous mathematician who revolutionized algebraic geometry came to our peaceful area for his retirement.

Sponsors of Sparse Days in St Girons III

The International Conference in St Girons is one of the main highlights of the CIMI semester. Sparse Days at St Girons III follows the model of earlier meetings held in St Girons in 1994 and 2003. Thanks to CIMI and the other sponsors, it is on a much grander scale than the Annual Sparse Days meeting hosted by CERFACS in Toulouse.

We would thus like to thank the many sponsors for their support for this meeting. A main theme of Sparse Days is the implementation of numerical methods on high performance and even extreme scale computers and we are very grateful to the computer vendors ATOS-Bull, IBM, and NVIDIA for their generous support acknowledged in the programme. CERFACS is supporting us as this replaces their 2015 Sparse Days and we are pleased by the support of the academic institutions from Toulouse: Université de Toulouse, INPT, IMT, CNRS, and IRIT, most of whom are additionally supporting us through the Centre International de Mathématiques et d'Informatique.

Finally, we would like to acknowledge the very strong support that we have had from the political sector: from the Région Midi-Pyrénées and from our hosts the Conseil Départemental de l'Ariège and the Ville de Saint Girons.

List of speakers to Sparse Days in St Girons III

Name	Institution	Country
ASHCRAFT Steve	Livermore Software Technology Corporation	USA
BISSELING Rob	Utrecht University	USA
CANNING Andrew	Lawrence Berkeley National Laboratory, UC Davis	USA
DARVE Eric	Stanford University	USA
DAVIS Tim	Texas A& M University	USA
DEMMEL Jim	UC Berkeley	USA
DOLBEAU Romain	ATOS	France
EATON Joe	NVIDIA	USA
FAVERGE Mathieu	INRIA Bordeaux	France
GANDER Martin	University of Geneva	Switzerland
HOGG Jonathan	STFC-RAL	England
HORAK David	University of Ostrava	Czech Republic
IL'IN Valery	Novosibirsk State University	Russia
KAYA Oguz	ENS Lyon	France
KNIGHT Philip	University of Strathclyde	Scotland
LANGOU Julien	University of Colorado Denver	USA
LI Xiaoye	Lawrence Berkeley National Laboratory	USA
LOPEZ Florent	University of Toulouse	France
MARTINSSON Gunnar	University of Colorado Denver	USA
MATSUOKA Satoshi	Tokyo Institute of Technology	Japan
MEURANT Gérard	CEA Paris	France
MICHEL Loïc	SNCF	France
NAKAJIMA Kengo	University of Tokyo	Japan
NESTEROV Yuri	Université Catholique de Louvain	Belgium
NG Esmond	Lawrence Berkeley National Laboratory	USA
NOTAY Yvan	Université Libre de Bruxelles	Belgium
POTHEN Alex	Purdue University	USA
PUDOV Sergey	INTEL MKL Team	Russia
RENNICH Steven	NVIDIA	USA
ROMMES Joost	Mentor Graphics	USA
SAMEH Ahmed	Purdue University	USA
SCHILDERS Wil	TU Eindhoven	Netherlands
SCOTT Jennifer	STFC-RAL	England
SOLOVYEV Sergey	Novosibirsk State University	Russia
SORENSEN Dan	Rice University	USA
TOINT Philippe	University Namur	Belgium
TITLEY-PELOQUIN David	McGill University	Canada
TŮMA Miroslav	Czech Academy of Sciences	Czech Republic
UÇAR Bora	ENS Lyon	France
VEZOLLE Pascal	IBM	USA
VICENTE Luis Nunes	University of Coimbra	Portugal
XIA Jianlin	Purdue University	USA
ZHANG Yin	Rice University	USA

Programme

	Sunday, June 28th	
19:00		Registration at L'Auberge d'Antan Welcome reception and dinner at l'Auberge d'Antan Sponsored by local authorities
All talks in Salle Max Linder (Place Jean Ibanes), St Girons		
Next to the Hotel de Ville (Town Hall)		
Lunches in ground floor of Town Hall		
Posters	in lunch room	
	L. Michel	Derivative-free optimization by model-free control approach
	D. Titley-Peloquin	The componentwise symmetric backward error for sparse symmetric systems of equations
	Monday, June 29th	
8:30		Registration
8:55		Welcome by the Mayor of St Girons
9:00	J. Demmel	Communication avoiding algorithms for linear algebra and beyond
9:40	E. Ng	An update on greedy ordering heuristics
10:10	J. Langou	Improving the communication lower bounds for matrix-matrix multiplication
10:40		Coffee break
11:10	A. Pothen	The Virtual Scalpel: real-time matrix computations and finite elements for surgery
11:40	M. Faverge	Blocking strategy optimizations for sparse direct linear solver on heterogeneous architectures.
12:10		Lunch (with contribution from ATOS)
14:10	T. Davis	Sparse SVD, and a GPU-accelerated sparse QR
14:40	J. Scott	Incomplete factorization preconditioners for large sparse least squares problems
15:10	F. Lopez	Task-based multifrontal QR solver for GPU-accelerated multicore architectures
15:30	M. Tůma	Algebraic preconditioning of symmetric indefinite systems
16:00		Coffee break
16:30	J. Eaton	Programming techniques for sparse linear algebra on GPUs
17:00	R. Dolbeau	Exploiting new CPU architectural features for sparse algorithms
17:30	P. Vezolle	OpenPower and IBM Exascale hybrid architecture overview and benefits for sparse matrices
20:00		Dinner at "Le Forail", Chemin de Pégoumas (St Girons), sponsored by ATOS
	Tuesday, June 30th	
9:00	R. Bisseling	Edge-based graph partitioning
9:30	P. Knight	Using matrix scaling to identify block structure
10:00	B. Uçar	Two approximation algorithms for bipartite matching on multicore architectures
10:30		Coffee Break
11:00	S. Rennich	GPU acceleration of sparse Cholesky factorization
11:30	J. Hogg	Practical manycore pivoting
12:00	S. G. Pudov	Intel Math Kernel Library Inspector-Executor Sparse BLAS API for iterative computations
12:30		Lunch (with contribution from IBM)
14:20	O. Kaya	Scalable sparse tensor decompositions in distributed memory systems
14:40	K. Nakajima	Parallel multigrid solvers on multicore cluster architectures
15:10	S. Solovyev	Parallel implementation of multifrontal hierarchically semi-separable solver for 3D Helmholtz problem
15:40		Coffee break
16:10	L. N. Vicente	Direct search based on deterministic and probabilistic descent
16:40	P. Toint	Evaluation complexity in constrained and unconstrained nonlinear optimization
17:10	Y. Nesterov	Complexity bounds for primal-dual methods minimizing the model of objective function
19:30		Gala Dinner at "Le domaine du palais", St Lizier

Programme

		Wednesday, July 1st
8:50	S. Matsuoka	Harnessing the deep memory hierarchy with communication reducing algorithms for Exascale
9:30	Y. Notay	Multilevel solution of linear systems with graph Laplacian matrices
10:00	D. Horak	PERMON toolbox
10:30		Coffee break
11:00	J. Rommes	Challenges in numerical simulation of electrical circuits
11:30	D. Sorensen	A DEIM induced CUR factorization
12:00		Lunch (with contribution from NVIDIA)
14:00	C. Ashcraft	Separability, partitions and coverings
14:30	E. Darve	Fast multipole method and hierarchical matrices
15:00	S. Li	Hierarchically parallel algorithms for linear systems with data-sparse (sub)matrices
15:30	G. Martinsson	Accelerating direct solvers using randomized methods
16:00		Coffee break
16:30	W. Schilders	Model order reduction for coupled problems using low rank approximations
17:00	A. Canning	Sparse matrix problems in first-principles electronic structure methods for materials science and chemistry
17:30	V. P. Il'In	Parallel approaches of domain decomposition methods on the basis of low-rank matrix approximation
19:00		Aperitif offered by shopkeepers of St Girons.
		Thursday, July 2nd
8:50	M. Gander	Five decades of time parallel time integration
9:30	A. Sameh	A scalable parallel algorithm for large sparse symmetric eigenvalue problems
10:00	G. Meurant	On the convergence of the Arnoldi process for computing eigenvalues
10:30		Coffee break
10:50	Y. Zhang	Block algorithms in an augmented Rayleigh-Ritz framework for large-scale exterior eigenpair computation
11:20	J. Xia	Superfast direct eigensolvers for large symmetric matrices and the structured perturbation analysis
11:50		Farewell by Mayor of St Girons
11:55		Lunch
13:00		Buses leave for Toulouse

The organizers thank the EU EESI project for their active participation and support for the conference.

Talks

Cleve Ashcraft *Separability, partitions and coverings*

The past two decades have seen an abundance of "data sparse" methods to represent and compute with a sparse or dense matrix. Storage and operation counts can be much better than using a direct factor and solve approach. The alphabet is well exercised, with BLR, H, HBS, HOLDR, HSS and H2 methods.

There is a common characteristic — at some point the matrix of interest either has, or is equivalent to, the following representation.

$$A_{\mathcal{I},\mathcal{I}} = D_{\mathcal{I},\mathcal{I}} + U_{\mathcal{I},\alpha} B_{\alpha,\beta} V_{\mathcal{I},\beta}^T$$

If $A_{\mathcal{I},\mathcal{I}}$ is nonsingular, then $D_{\mathcal{I},\mathcal{I}}$ is nonsingular. Think of $D_{\mathcal{I},\mathcal{I}}$ as the near field, and $H_{\mathcal{I},\mathcal{I}} = U_{\mathcal{I},\alpha} B_{\alpha,\beta} V_{\mathcal{I},\beta}^T$ as the farfield. The columns of $U_{\mathcal{I},\alpha}$ are the left modes, the columns of $V_{\mathcal{I},\beta}$ are the right modes. We would like $U_{\mathcal{I},\alpha}$ and $V_{\mathcal{I},\beta}$ to be orthonormal, for then $B_{\alpha,\beta}$ acts like a weight matrix.

The BLR, H and HODLR methods construct $U_{\mathcal{I},\alpha}$ in an additive sense. They *partition* the matrix into submatrices, i.e., the submatrices have disjoint nonzero structure. One builds up the $U_{\mathcal{I},\alpha}$ matrix by block columns, one block column for each submatrix in the partition. The $U_{\mathcal{I},\alpha}$ matrix is sparse. Each $U_{\mathcal{I},\alpha_i}$ and $V_{\mathcal{I},\beta_i}$ has a strict subset of rows that are nonzero. Each $U_{\mathcal{I},\alpha_i}$ is orthonormal, but $U_{\mathcal{I},\alpha}$ need not be orthonormal. $B_{\alpha,\beta}$ is block diagonal.

A second approach is to construct a separable representation recursively. We start with $A_{\mathcal{I},\mathcal{I}} = D_{\mathcal{I},\mathcal{I}} + U_{\mathcal{I},\alpha_0} B_{\alpha_0,\beta_0} V_{\mathcal{I},\beta_0}^T$. Assume B_{α_0,β_0} has the same form.

$$B_{\alpha_0,\beta_0} = D_{\alpha_0,\beta_0} + U_{\alpha_0,\alpha_1} B_{\alpha_1,\beta_1} V_{\beta_0,\beta_1}^T$$

Continue in this manner. The $U_{\mathcal{I},\alpha}$ matrix is made from block columns $U_{\mathcal{I},\alpha_i}$ that have an implicit structure.

$$U_{\mathcal{I},\alpha_i} = U_{\mathcal{I},\alpha_{i-1}} U_{\alpha_{i-1},\alpha_i} \text{ for } i > 0$$

The $\alpha_0, \alpha_1, \ldots \alpha_k$ modes on the levels form a *covering* of α, not a partition. The implicit representations of $U_{\mathcal{I},\alpha_i}$ and $V_{\mathcal{I},\beta_i}$ mean reduced storage and operation count over their additive representation. When each $U_{\mathcal{I},\alpha_i}$ is block diagonal, and $U_{\mathcal{I},\alpha_i}$ nests with $U_{\mathcal{I},\alpha_{i+1}}$, we have the tree-based H^2, HBS and HSS methods.

Rob Bisseling *Edge-based graph partitioning*

Partitioning the edges of a graph for the purpose of parallel graph computations is closely related to two-dimensional partitioning of the nonzeros of a sparse matrix for the purpose of parallel sparse matrix–vector multiplication. In previous work, graph matching algorithms with an edge-centric data distribution have been shown to incur less communication and to balance the work load better than those with a vertex-centric distribution.

In this talk, we present the recent medium-grain method for 2D sparse matrix partitioning based on a simple initial split of an $m \times n$ matrix A followed by a multilevel 1D partitioning of an $(m+n) \times (m+n)$ matrix B. This method is fast and leads to lower communication volumes than previous methods.

Furthermore, we present an exact partitioner MondriaanOpt, which optimally bipartitions small sparse matrices in a branch-and-bound fashion. We demonstrate the use of MondriaanOpt by giving optimal communication volumes and showing pictures of optimal bipartitionings for a set of matrices from the University of Florida collection with up to 129,042 nonzeros. These results can serve as a test suite to benchmark the quality of heuristic solvers.

Joint work with Daniël Pelt (CWI, Amsterdam)

Andrew Canning *Sparse matrix problems in first-principles electronic structure methods for materials science and chemistry.*

First-principles materials science and Chemistry codes based on density functional theory (DFT) have become the largest user (by method) of computer cycles at scientific computer centers around the world. At NERSC (National Energy Research Scientific Computing Center) an estimated 20total cycles are used by DFT based codes. These codes are all solving the non-linear eigenfunction problem for the density functional theory based approximation to the many-body Schrodinger Equation (usually the Kohn-Sham form). While the computational cost of the standard dense matrix formalism scales as the cube of the number of atoms in the system in recent years new methods have been developed that exploit some form of sparsity in the eigenfunction problem to reduce the scaling to linear or squared in the number of atoms. In this talk I will give an overview of the different methods and the sparse eigenvalue problems arising in these methods. In particular I will focus on methods that explicitly work with spatially localized eigenfunctions.

Eric Darve *Fast multipole method and hierarchical matrices*

We will present joint-work between Stanford and INRIA on the use of hierarchical matrices for fast algebra. Hierarchical matrices is a ubiquitous format that can be used to perform fast matrix-vector products, factor dense matrices, eliminate fill-in in sparse matrix algebra, and for hybrid solvers with domain decomposition. We will present recent progress on these methods. These methods have been released as part of several software: ScalFmm, PaStiX, MaPHyS, Dense_HODLR, and Sparse_MultiFrontal.

Joint work with Pierre Ramet, Luc Giraud, Jean Roman, and Olivier Coulaud.

Tim Davis *Sparse SVD, and a GPU-accelerated sparse QR*

Two related sparse direct methods are presented: sparse QR factorization on the GPU, and a sparsity-preserving reduction to bidiagonal form. Our multifrontal GPU-accelerated sparse QR relies on a novel 'bucket scheduler' algorithm that enables the GPU(s) to exploit a high degree of irregular parallelism. Entire subtrees of the etree reside on the GPU, and the GPU factorizes the frontal matrices in parallel, performing the frontal matrix assembly and factorization itself, without having the data transfered back and forth between the GPU and CPU. For a single subtree, the CPU transfers to the GPU a sparse submatrix of matrix to be factorized, and receives back the resulting factored matrix.

This sparse QR is useful in its own right, for many applications, but it also forms the first step in a sparsity preserving SVD. The SVD starts with a direct method that reduces the sparse R (from QR) from skyline to bidiagonal form, via selective Givens rotations and block-Givens rotations that preserve the skyline.

Joint work with Sanjay Ranka, Nuri Yeralan, Wissam Sid-Lakhdar (sparse QR), and Siva Rajamanikam (sparse SVD).

Jim Demmel *Communication avoiding algorithms for linear algebra and beyond.*

Algorithms have two costs: arithmetic and communication, i.e. moving data between levels of a memory hierarchy or processors over a network. Communication costs (measured in time or energy per operation) already greatly exceed arithmetic costs, and the gap is growing over time following technological trends. Thus our goal is to design algorithms that minimize communication. We present algorithms that attain provable lower bounds on communication, and show large speedups compared to their conventional counterparts. These algorithms are for direct and iterative linear algebra, for dense and sparse matrices, as well as direct n-body simulations. Several of these algorithms exhibit perfect strong scaling, in both time and energy: run time for a fixed problem size drops proportionally to the number of processors p, with constant energy. Finally, we describe extensions to very general algorithms involving arbitrary loop nests and array accesses, in a way that may be incorporated into compilers.

Romain Dolbeau *Exploiting new CPU architectural features for sparse algorithms*

With the prophesied end of Moore's Law, compute architectures are changing. After evoking some possible consequences of those changes, we'll look to the short-term with the upcoming Intel Knights Landing processor. Scatter-gather, large vector, masking and stacked memory need all be understood to efficiently exploit the potentially complex memory structures and access patterns of sparse linear algebra algorithms.

Joe Eaton *Programming techniques for sparse linear algebra on GPUs.*

We discuss our efforts to provide a standardized C-level interface for common linear algebra operations on GPUs. Accelerating sparse linear algebra on the latest GPU architectures has real potential for performance gains of hundreds of percent over carefully tuned multi-core CPU-only implementations, but at what cost in complexity? This talk will address the programming approaches needed to utilize GPUs at scale for today's most challenging problems, and give a glimpse of the path forward to exascale applications in CFD, databases and search.

Mathieu Faverge *Blocking strategy optimizations for sparse direct linear solver on heterogeneous architectures.*

In the context of solving sparse linear systems, the nested dissection process partitions the matrix graph to minimize both the fill-in and the computational cost. We found that the classic Reverse Cuthill McKee algorithm used to order unknowns in supernodes might be enhanced to reduce the number of off-diagonal blocks by increasing their sizes. This turns into the same complexity for the factorization algorithm, but allows for more efficient BLAS kernels. On the other side, one might want to split the larger supernode to introduce more parallelism. The regular splitting strategy when applied locally impacts significantly the number of off-diagonal blocks and might have negative effect on the efficiency. In this talk, we present both a new strategy to improve supernodes ordering and splitting strategy that both enlarge the average off-diagonal block sizes without changing the computational cost of the factorization. Performance improvement gains on the supernodal solver PaStiX are shown on multi-cores and heterogeneous architectures.

Joint work with Grégoire Pichon, Pierre Ramet, and Jean Roman.

Martin J. Gander *Five decades of time parallel time integration.*

Time parallel time integration methods have received renewed interest over the last decade because of the advent of massively parallel computers, which is mainly due to the clock speed limit reached on today's processors. When solving time dependent partial differential equations, the time direction is usually not used for parallelization. But when parallelization in space saturates, the time direction offers itself as a further direction for parallelization. The time direction is however special, and for evolution problems there is a causality principle: the solution later in time is affected (it is even determined) by the solution earlier in time, but not the other way round. Algorithms trying to use the time direction for parallelization must therefore be special, and take this very different property of the time dimension into account.

I will show in this talk how time domain decomposition methods were invented, and give an overview of the existing techniques. Time parallel methods can be classified into four different groups: methods based on multiple shooting, methods based on domain decomposition and waveform relaxation, space-time multigrid methods and direct time parallel methods. I will show for each of these techniques the main inventions over time by choosing specific publications and explaining the core ideas of the authors. This talk is for people who want to quickly gain an overview of the exciting and rapidly developing area of research on time parallel methods.

Jonathan Hogg *Practical manycore pivoting*

Strong scaling of sparse linear solvers on manycore architectures requires a solution to the latency bottleneck of traditional threshold partial pivoting. We describe our experiments using new pivoting heuristics on difficult real world problems. These include both multilevel a posteriori ("try-it-and-see") approaches that seek to avoid pivoting and communication avoiding variants that work with a representative subset of the data.

David Horak *PERMON toolbox*

We shall present our new software called PERMON (Parallel, Efficient, Robust, Modular, Object-oriented, Numerical) toolbox [3]. It is based on PETSc and it combines domain decomposition methods especially of FETI type (Total-FETI) [1] and quadratic programming (QP) algorithms (such as MPRGP - Modified Proportioning with Reduced Gradient Projections or SMALBE - SemiMonotonic Augmented Lagrangian algorithm for Bound and Equality constraints), both developed by Dostal [2]. The algorithms have the rate of convergence in terms of the bounds on the spectrum of the Hessian matrices. This combination enjoys both numerical and parallel scalabilities for the solution of the contact problems of elasticity.

The core solver layer consists of several modules: PermonQP for unconstrained and equality-constrained QP, its PermonFLLOP extension for FETI, and PermonIneq extension providing algorithms for inequality-constrained QP. Other modules include PermonPlasticity for plasticity, PermonImage for the image registration, PermonMultiBody for particle dynamics, and others. The main idea of PermonQP is separation of problems and solvers. A QP transformation derives a new QP from the given QP, thus, sort of doubly linked list is generated where every node is a QP.

In case of linear problems, we only need to solve the coarse problem (CP) which is part of the projector in the Hessian. This action is implemented in all parallel sparse direct solvers as a "solve" function which comprises of the forward elimination step followed by the backward elimination step in turn. In case of contact problems, we need a stand-alone transformation matrix denoting a regular matrix that defines the orthonormalization of the rows of coarse space matrix to be able to premultiply by that the RHS vector in dual equality constraint before homogenization or to update SMALSE RHS vector and multiplier. Unfortunately, parallel sparse direct solvers like MUMPS typically use its efficient inner representation of factors, optimized for parallel forward/backward elimination, and they do not typically provide to user neither a stand-alone factor nor separate forward and backward eliminations. So, another approach avoiding to the coarse space matrix orthonormalization is proposed. It is based on the stiffness matrix null-space orthonormalization and constraint matrix scaling by diagonal matrix that stores in each diagonal entry the multiplicity of the corresponding dof. Numerical experiments demonstrating the performance including the highlights with model and engineering problems will be presented at the conference.

Joint work with Zdenek Dostal, Vaclav Hapla, Lukas Pospisil, Alexandros Markopoulos, Martin Cermak, Alena Vasatova, and Radim Sojka, all from IT4I.

Valery P. Il'in *Parallel approaches of domain decomposition methods on the basis of low-rank matrix approximation.*

The efficiency and performance of two-level parallel iterative processes in the Krylov subspaces are investigated in solving large sparse symmetric and non-symmetric systems of linear algebraic equations arising from grid approximations of multi-dimensional boundary value problems for PDEs. A special attention is being given to optimization of the subdomain overlapping size, to the types of interface conditions on adjacent boundaries in the domain decomposition method, and to the aggregation (or coarse grid correction) algorithms. An acceleration of the iterations over the subdomains is based on preconditioning matrices which are low-rank approximations of the original matrices. Various approaches are used for different orders of basic interpolation functions and for different positions of the coarse grid nodes.

An outer iterative process is based on the additive Swartz algorithm, while a parallel solution of subdomain algebraic systems is affected by the direct or the preconditioned iterative Krylov method. A crucial point in implementation of these approaches is a technology of the balanced forming the extended algebraic subsystems in the overlapping subdomains. A comparative analysis of the influence of various parameters is carried out for a representative set of numerical experiments.

Some issues related to the scalability of parallelization on multi-processor systems with distributed and shared memory are discussed.

Joint work with Yana L. Gurieva (Institute of Computational Mathematics)

Oguz Kaya *Scalable sparse tensor decompositions in distributed memory systems*

We investigate an efficient parallelization of the most common iterative sparse tensor decomposition algorithms on distributed memory systems. A key operation in each iteration of these algorithms is the matricized tensor times Khatri-Rao product (MTTKRP). This operation amounts to element-wise vector multiplication and reduction depending on the sparsity of tensor. We investigate a fine and a coarse-grain task definition for this operation, and propose hypergraph partitioning-based methods for these task definitions to achieve load balance as well as reduce communication requirements. We also design a distributed memory sparse tensor library, HyperTensor, which implements a well-known algorithm for the CANDECOMP/PARAFAC (CP) tensor decomposition using the task definitions and the associated partitioning methods. We use this library to test the proposed implementation of MTTKRP in CP decomposition context, and report scalability results up to 1024 MPI ranks. We demonstrate up to 194 fold speedups using 512 MPI processes on a well-known real world data, and significantly better performance results with respect to a state of the art implementation.

Joint work with Bora Uçar, CNRS and ENS Lyon.

Philip Knight *Using matrix scaling to identify block structure*

We can apply a two-sided diagonal scaling to a nonnegative matrix to render it into doubly stochastic form if and only if the matrix is fully indecomposable. The scaling often reveals key structural properties of the matrix as the effects of element size and connectivity are balanced. Exploiting key spectral properties of doubly stochastic matrices, we will show how to use the scaling to reveal hidden block structure in matrices without any prior knowledge of the number and size of the blocks. In particular, the structure of the basis of the principal singular vectors of the scaled matrix allows us to perform a multi-way partition of the rows and columns of the matrix in the spirit of the Fiedler vector. The application of a Canny filter to one or more singular vectors allows us to detect the number of clusters automatically .

In order to be able to apply our method to as wide a class of problems as possible we introduce pre-processing steps to guard against finding spurious clusters of elements and to make sure our partitions are of a reasonable size.

The method can be thought of as a bi-clustering and we compare our new approach against existing bi-clustering algorithms.

This work was performed in collaboration with Iain Duff, Sandrine Mouysset, Daniel Ruiz and Bora Ucar.

Julien Langou *Improving the communication lower bounds for matrix-matrix multiplication*

We consider communication lower bounds for matrix-matrix multiplication in the sequential case (number of messages and total volume of messages). Our new proof technique improves the known lower bound for the number of reads and writes in the sequential memory case. This lower bound can be adapted to hierarchical memory and parallel distributed memory cases.

Xiaoye S. Li *Hierarchically parallel algorithms for linear systems with data-sparse (sub)matrices*

It was long discovered that for structured matrices, such as Hilbert matrix, Toeplitz matrix and the matrices from the BEM methods for integral equations, certain off-diagonal submatrices are numerically rank deficient. Many compression techniques (e.g., truncated SVD) have been developed to exploit low-rankness. These compact forms can be used to design various asymptotically faster and memory-efficient linear algebra algorithms, beyond matrix-vector multiplication. In particular, we have pioneered the work on using hierarchically semi-separable (HSS) compression to develop scalable factorization and solution algorithms for direct solvers or preconditioners.

In this talk, we will briefly survey a number of compression techniques. We will illustrate how the HSS structure can be employed to speed up the solution process for the data-sparse matrices in dense form, as well as for the data-sparse submatrices in sparse form. We will show that both in theory and in practice, the HSS-structured factorizations have much lower complexity than the traditional factorization algorithm. The performance results on the parallel machines will be shown with our recent open-source library STRUMPACK (`http://portal.nersc.gov/project/sparse/strumpack/`).

Joint work with A. Napov, P. Ghysele and F.-H. Rouet.

Florent Lopez *Task-based multifrontal QR solver for GPU-accelerated multicore architectures*

Recent studies have shown the potential of task-based programming paradigms for implementing robust, scalable sparse direct solvers for modern computing platforms. Yet, designing task flows that efficiently exploit heterogeneous architectures remains highly challenging. In this talk we first discuss the data partitioning using a method suited to heterogeneous platforms allowing task granularity to be sufficiently large to obtain a good acceleration factor on GPU but capable of generating enough parallelism in the task graph. Secondly we handle the task scheduling with a strategy capable of taking into account workload and architecture heterogeneity at a reduced cost. Finally we propose an original evaluation of the performance obtained in our solver on a test set of matrices.

Gunnar Martinsson *Accelerating direct solvers using randomized methods*

The last decade has seen rapid progress in constructing direct (as opposed to iterative) solvers for the linear systems arising upon the discretization of elliptic PDEs. Solvers with linear, or close to linear, complexity have been described for both sparse systems associated with finite element or finite difference discretizations, and for the dense systems associated with the discretization of the corresponding integral equations.

The talk will describe how randomized sampling can be used to accelerate many of the linear algebraic computations that form the backbone of most direct solvers. Randomized schemes can very advantageously be used to compute approximate low rank approximations to matrices. They can also be used to construct so called "data sparse" representations of dense matrices whose off-diagonal blocks are of numerically low rank, such as, e.g., H-matrices, or Hierarchically Semi-Separable matrices.

Satoshi Matsuoka *Harnessing the deep memory hierarchy with communication reducing algorithms for Exascale.*

Future exascale systems is predicted to be largely constrained by memory bandwidth and capacity, rather than FLOPS, in the context of real application performance. While it is plausible to accommodate either of the bandwidth or capacity, it would be difficult to satisfy both at the same time due to power and cost constraints. Although the new breed of non-volatile memory promises to provide lower power and higher density, due to their access characteristics they will likely supplement, not replace, DRAMs and associated caches in the processors. As a result, memory hierarchy will become increasingly deeper, and unless the algorithms will cope with them explicitly, it would be difficult to attain efficient performance.

In order to satisfy both bandwidth and capacity requirements simultaneously, we are working on several breed of communication-reducing algorithms tailored to cope with the deep memory hierarchy. These include (1) partitioned variant of the recently proposed Cell-C-Sigma for sparse matrix operation on GPUs, (2) effective multi-level temporal blocking with programmability assist, and (3) offloading of sparse graph structures to NVM in graph algorithms. The speedup results are quite dramatic for each one, allowing large capacity problems to be resident in lower tiers of the memory hierarchy while preserving performance, without specialized hardware.

Gérard Meurant *On the convergence of the Arnoldi process for computing eigenvalues*

In this talk we are interested in the computation of eigenvalues with the Arnoldi process for a diagonalizable matrix. We consider finding closed-form expressions for the coefficients of a polynomial whose roots are the Ritz values. They are given in terms of the eigenvalues and eigenvectors of A as well as the starting vector. This also yields an expression for the coefficients of the FOM residual polynomial. Then we do the same for the harmonic Ritz values and we obtain an expression for the coefficients of the GMRES residual polynomial. The knowledge of the coefficients of these polynomials can be used to obtain bounds for the distances of the roots to eigenvalues of A.

Kengo Nakajima *Parallel multigrid solvers on multicore cluster architectures*

Parallel multigrid method is expected to be a powerful tool for large-scale computations, but includes both of serial and parallel communication processes which are generally expensive. The serial communication is the data transfers through memory hierarchies of each processor, while the parallel one is by message passing between computing nodes using MPI. This presentation summarizes recent efforts of optimization of serial and parallel communications in parallel MGCG (conjugate gradient with multigrid preconditioning) solvers with geometric multigrid procedures using up to 4,096 nodes (65,536 cores) of Fujitsu PRIMEHPC FX10 [1]. Performance of both of flat MPI and HB M $\times$ N (M: number of threads on each MPI process, N: number of MPI processes on each node) has been evaluated. In the present work, new format for sparse matrix storage based on sliced ELL, which has been well-utilized for optimization of SpMV, is proposed for optimization of serial communication on memories, and hierarchical coarse grid aggregation (hCGA) is introduced for optimization of parallel communication by message passing. The parallel MGCG solver using the sliced ELL format provided performance improvement in both weak scaling (25%-31%) and strong scaling (9%-22%) compared to the code using the original ELL format. Moreover, hCGA provided excellent performance improvement in both weak scaling (1.61 times) and strong scaling (6.27 times) for flat MPI parallel programming model [1]. hCGA was also effective for improvement of parallel communications. Computational amount of coarse grid solver for each core of flat MPI is 256 (=16x16) times as large as that of HB 16x1. Therefore, hCGA is expected to be really effective for HB 16×1 with more than 2.50×10^5 nodes of Fujitsu FX10, where the peak performance is more than 60 PFLOPS. CGA and hCGA include a various types of parameters, and the optimum values of those were derived through empirical studies in the present work. Development of methods for automatic selection of these parameters is also an interesting technical issue for future work. Optimum parameters can be estimated based on calculation of computational amounts, performance models, parameters of hardware, and some measured performance of the system. But it is not so straightforward. Because some of these parameters also make effects on convergence, construction of such methods for automatic selection is really challenging.

[1] Nakajima, K., Optimization of Serial and Parallel Communications for Parallel Geometric Multigrid Method, Proceedings of the 20th IEEE International Conference for Parallel and Distributed Systems (ICPADS 2014) (2014) 25-32

Yuri Nesterov *Complexity bounds for primal-dual methods minimizing the model of objective function*

We provide Frank-Wolfe ($\equiv$ Conditional Gradients) method with a convergence analysis allowing to approach a primal-dual solution of convex optimization problem with composite objective function. Additional properties of complementary part of the objective (strong convexity) significantly accelerate the scheme. We also justify a new variant of this method, which can be seen as a trust-region scheme applying the linear model of objective function. Our analysis works

also for a quadratic model, allowing to justify the global rate of convergence for a new second-order method. To the best of our knowledge, this is the first trust-region scheme supported by the worst-case complexity bound.

Esmond G. Ng *An update on greedy ordering heuristics*

The minimum degree (MD) and minimum local fill (MLF) algorithms are two often mentioned bottom-up heuristics for reordering the rows and columns of a sparse symmetric matrix to reduce the cost of computing its symmetric factorization. Between the two, the MD algorithm is probably the best-known and most widely-used heuristic. The primary reason why the MLF algorithm has not been as popular as the MD algorithm is because of the general belief that a MLF ordering is very expensive to compute. We have recently described an efficient implementation of the MLF algorithm using an updating scheme due to Wing and Huang. In this talk, we will report on an extensive experiment, in which several implementations based on the MD and the MLF heuristics were compared with an implementation of the nested dissection algorithm. As far as we know, this may be the most exhaustive comparision that has ever been conducted.

Joint work with Barry W. Peyton (Dalton State College)

Yvan Notay *Multilevel solution of linear systems with graph Laplacian matrices.*

Graphs appear in a large variety of applications. As a consequence, their structure is very diverse, making difficult the design of a general purpose solver for the linear systems associated with graph Laplacian matrices. Direct solvers can be occasionally fast, and simple iterative methods such as the Gauss–Seidel method are efficient in a number of cases. However, more sophisticated approaches are needed to cope with large graphs without much a priori knowledge on their structure or their algebraic properties.

Among these, it is natural to consider multilevel algorithms based on the aggregation of the unknowns. Indeed, on the one hand, aggregation is a standard process to obtain a coarse representation of a graph. On the other hand, related methods have been proved efficient to solve discrete scalar elliptic PDEs [1,2], which includes the Laplacian of mesh-type graphs.

Now, performing experiments on graphs from the University of Florida Collection, it turns out that, whereas existing methods and simple (greedy type) aggregation schemes are often efficient, they cannot be considered as robust: depending on the case at hand, they can exhibit slow convergence, or excessive complexity (because the size of the graph does not decrease fast enough from one level to the next), or both.

This led us to develop a new aggregation method, which combines a greedy type scheme (forming sometimes aggregates with hundreds of unknowns) with a quality control along the lines of the approach already followed in [2], aiming at guaranteeing that the two-level condition number is below a prescribed threshold.

Experiments show that the resulting method solves robustly *all* graphs from the University of Florida and some other collections, and is on average significantly faster than any competitor.

[1] Y. Notay, An aggregation-based algebraic multigrid method, Electron. Trans. Numer. Anal., 37 (2010), pp. 123-146.

[2] A. Napov and Y. Notay, An algebraic multigrid method with guaranteed convergence rate, SIAM J. Sci. Comput., 34 (2012), pp. A1079-A1109.

Joint work with Artem Napov.

Alex Pothen *The Virtual Scalpel: real-time matrix computations and finite elements for surgery*

Eye surgeons learn their craft on animal models, but these models differ from human eyes in their viscoelastic properties. Simulators based on haptics and computer graphics have the potential to provide better training tools for surgery. These simulators need to update images of the eye from ten to hundred times per second to provide realistic visualization.

We present fast matrix computation algorithms to support interactive visualization of solid finite element models of human organs. Integrating support for cutting with real-time finite element solution methods is a computational challenge due to two reasons: First, high update rates are required for graphical and haptic rendering; Second, the connectivity changes due to the cutting (removal and insertion of nodes and elements, and re-meshing around the cut) necessitate corresponding changes to the underlying system of linear equations. We describe an algorithm that provides a fast visualization functionality through an augmented matrix approach, a hybrid linear equation solver, and sophisticated exploitation of sparsity in the matrices and vectors. The time complexity of our algorithm is bounded by a term linear in the number of nonzeros in the Choleslky factors of the initial matrix. We are able to provide ten to hundreds of updates per second for meshes with hundreds of thousands of elements. To the best of our knowledge, this is the first such result for meshes of this size.

The fast update problem arises in contingency analysis of electric power grids also, when operators have to pre-compute strategies for actions to take when a specified number of the elements in the grid fail. This problem is computationally intensive due to the large number of cases that must be analyzed. The augmented matrix approach makes it possible to successively update the solutions with increasing numbers of failed elements.

Joint work with Yu-Hong Yeung (Purdue) and Jessica Crouch (Old Dominion University).

Sergey G Pudov *Intel Math Kernel Library Inspector-Executor Sparse BLAS API for iterative computations*

The implementation of Sparse BLAS functionality in the Intel Math Kernel Library (Intel MKL) versions not higher than 11.2 is based on the NIST* Sparse BLAS C implementation. This API uses a single function call for any compute operation and does not allow passing optimization information between function calls. This limits certain aggressive optimizations, such as balancing based on matrix sparsity patterns, matrix reordering, and even matrix format changes. These optimizations require time compared to one sparse-matrix vector multiplication and become beneficial only when multiple operations are performed with a single matrix, such as in iterative solvers. Intel MKL 11.3 Beta introduces an inspector-executor API, which uses a two-step approach to computations. The analysis stage is used to inspect the matrix sparsity pattern and apply matrix structure changes. The information from the analysis stage is used in subsequent calls to do computations with higher performance. The API offers a consistent support for C- and Fortran-style data layouts (row- and column-major) and indexing (zero-based and one-based), as well as combinations of these. It supports key sparse matrix storage formats: CSR (CSC), COO and BSR. I will discuss optimizations made to support iterative solvers with matrix-vector multiplications and triangular solvers aimed to achieve scalability on Intel Xeon and Intel Xeon Phi processors.

Steven C. Rennich *GPU acceleration of sparse Cholesky factorization.*

Sparse matrix factorization, a critical algorithm in many science, engineering and optimization applications, has had difficulty effectively leveraging a large portion of the computational resources offered by GPUs both in single-node and cluster environments. Reasons for this difficulty include the inherent irregularity of data and computation in a sparse direct factorization, slow PCIe communication between the GPU can the CPU, and the complexity of writing and debugging massively parallel code.

To better understand these issues and discover / demonstrate how GPUs can be applied to the sparse factorization operation, we have been engaged in two projects. A detailed GPU optimization of the left-looking, supernodal Cholesky factorization method in SuiteSparse/CHOLMOD, and a minimally-invasive optimization of the right-looking, multi-frontal Cholesky method in the Watson Sparse Matrix Package (WSMP).

In this talk we will discuss the optimization strategies followed, performance benefits achieved, current performance limiters and avenues for further improvement for both WSMP and SuiteSparse/CHOLMOD.

Joint work with Timothy A. Davis (Texas A&M), Anshul Gupta (IBM), Natalia Gimelshein (NVIDIA), and Send Koric (University of Illinois / NCSA).

Joost Rommes *Challenges in numerical simulation of electrical circuits*

Solution of large systems of (non)linear equations is among the most computationally intensive tasks in numerical simulation of electrical circuits. A wide range of methods has been developed by experts in electrical engineering and scientific computing to meet the industry's requirements on speed and accuracy, but successfully combining these methods is a major challenge for several reasons, including the scattering over various libraries, software architectural decisions, and unknown and/or hardly predictable effects on accuracy when combining methods. In this presentation we give an impression of the algorithmic and architectural choices that have to be made, with main emphasis on trade-offs between flat and hierarchical solvers, iterative and direct solvers, exact and inexact methods, and full and reduced order models.

Joint work with Olivier Maury

Ahmed H. Sameh *A scalable parallel algorithm for large sparse symmetric eigenvalue problems.*

The trace-minimization scheme (TraceMin) proved to be a robust and scalable parallel algorithm for obtaining few of the smallest eigenpairs of large sparse symmetric eigenvalue problems $Ax = \lambda Bx$, where A is symmetric and B is symmetric positive definite. Although the first version of this algorithm was published in 1982 (more than three decades ago) exhaustive tests have shown that TraceMin is quite competitive to many eigensolvers introduced since the late 1990s. In this presentation we outline simple modifications of TraceMin that allows computing large number of eigenpairs belonging to any intermediate interval in the spectrum.

Joint work with Alicia Klinvex (Purdue)

Wil Schilders *Model order reduction for coupled problems using low rank approximations.*

Model order reduction (MOR) and numerical linear algebra are fields that are very closely related, especially when Krylov subspace methods are used for the reduction. In addition, the systems to be reduced are also sparse. Hence, it is not surprising that the field has seen rapid development of effective methods, starting with the Pade-via-Lanczos (PVL) method by Freund and Feldmann in 1994. Many issues have been resolved, so that MOR for linear problems is well developed by now.

In recent years, the focus for model order reduction is shifting to parameterized problems, nonlinear problems, differential-algebraic equations and coupled systems. In this talk, we will discuss a method we recently developed for the latter class of problems. It makes use of low rank approximations of the sparse coupling blocks by performing generalized singular value decompositions (GSVD), enabling a splitting of the coupled problem into a number of sub problems for which conventional MOR techniques can be used. The GSVD ensures that a suitable weighting of the off diagonal blocks with respect to the diagonal blocks is used, thereby identifying strong and weak couplings automatically.

The method will be discussed in more detail, and examples originating in the electronics industry will be used to show the effectiveness.

Jennifer Scott *Incomplete factorization preconditioners for large sparse least squares problems.*

To solve large-scale sparse linear least squares problems iterative methods can potentially offer a number of important advantages over direct methods. However, for iterative methods to be effective, a preconditioner is required. In the last decade, a number of preconditioners have been proposed (including RIF, MIQR, BIF and BA-GMRES). In this talk, we report on using our recent limited memory incomplete factorization preconditioners to solve least squares problems. Our incomplete factorization is based on the ideas of Tismentsky and has been shown to be effective, efficient and robust for symmetric systems coming from a wide range of applications. We explain the approach and its possible use for least squares problems. Numerical results for test examples arising from practical problems are used to illustrate the strengths and limitations of the approach.

Sergey Solovyev *Parallel implementation of multifrontal hierarchically semi-separable solver for 3D Helmholtz problem.*

We present a parallel implementation of the multifrontal sparse solver to solve 3D Helmholtz problem. It is based on Gauss-elimination direct solver coupled with Nested Dissection reordering approach and compressing factors by using low-rank approximation technique and HSS format. After discretization of Helmholtz equation with perfectly matched layer (PML) boundary conditions arises the complex symmetric matrix which is factorized as LDL^T in double precision arithmetic. Low-rank approximation of the off-diagonal blocks is performed by cross approximation technique (CA); the diagonal blocks are presented in HSS-format. Despite this algorithm is faster than traditional direct solvers and uses significantly less memory, it still cannot solve large Helmholtz problems which appeared in geophysics. Therefore, the parallel version of the algorithm has been developed. It is dedicated for on distributed memory systems (clusters) using hybrid OpenMP and MPI parallelization. OpenMP parallelization is performed due to optimized BLAS and LAPACK functions implemented in Intel MKL. MPI parallelization is based on distributing nodes of elimination tree between cluster nodes. In the direct solver algorithms, distribution of LDL^T-factors between cluster nodes can be predicted before starting the factorization (after reordering step). So, the main problem is constructing a balanced node distribution to improve MPI-scalability. Static distribution can be inefficient for low-rank solvers since size estimations for compressed LDL^T-factors are quite inexact. We propose a dynamic elimination tree nodes distribution to handle with this problem. At first, we suppose that all factors are not compressed and we start distributing them between cluster nodes as much as possible. After factorization and compressing process these factors are collected on certain cluster nodes. The next step is distributing and factorizing other nodes of elimination tree. So, while performing factorization process there are two types of cluster nodes: the first ones save factors (factorized elimination tree nodes) and send them by request to other cluster nodes for computing Schur complement of the next matrix columns; the other group of cluster nodes perform factorization. As the factorization process continues, more cluster nodes are storing factorized data and less nodes perform factorization. Proposed scheme is memory effective and can solve problems provided that compressed LDL^T-factors fit total clusters RAM (CRAM). The MPI-scalability of proposed scheme depends on the size of factors relative to the size of CRAM and increases with increasing number of cluster nodes.

The research described is supported by CRDF grant RUE1-30034-NO-13 and RFBR grants 14-01-31340, 14-05-31222, 14-05-00049.

Dan C. Sorensen *A DEIM induced CUR factorization.*

I will present a CUR matrix factorization based upon the Discrete Empirical Interpolation Method (DEIM). A CUR factorization provides a low rank approximate factorization of a given matrix $\mathbf{A}$ of the form $\mathbf{A} \approx \mathbf{CUR}$ where $\mathbf{C}$ is a subset of the columns of $\mathbf{A}$ and $\mathbf{R}$ is a subset of the rows of $\mathbf{A}$. The matrix $\mathbf{U}$ is constructed so that $\mathbf{CUR}$ is a good approximation to $\mathbf{A}$. Assuming a low rank SVD $\mathbf{A} \approx \mathbf{VSW}^T$ is available, the DEIM points for $\mathbf{V}$ and $\mathbf{W}$ are used to select the matrices $\mathbf{C}$ and $\mathbf{R}$ respectively. This approximate factorization will satisfy $\|\mathbf{A} - \mathbf{CUR}\|_2 = \mathcal{O}(\sigma_{k+1})$, the first neglected singular value of $\mathbf{A}$ for a certain construction of $\mathbf{U}$.

Philippe Toint *Evaluation complexity in constrained and unconstrained nonlinear optimization.*

We review the available results on the evaluation complexity of algorithms using Lipschitz-continuous Hessians for the approximate solution of nonlinear and potentially nonconvex optimization problems. Here, evaluation complexity is a bound on the largest number of problem functions (objective, constraints) and derivatives evaluations that are needed before an approximate first-order critical point of the problem is guaranteed to be found. We start by considering the unconstrained case and examine classical methods (such as Newton's method) and the more recent ARC2 method, which we show is optimal under reasonable assumptions. We then turn to constrained problems and analyze the case of convex constraints first, showing that a suitable adaptation ARC2CC of the ARC2 approach also possesses remarkable complexity properties. We finally extend the results obtained in simpler settings to the general equality and inequality constrained nonlinear optimization problem by constructing a suitable ARC2GC algorithm whose evaluation complexity also exhibits the same remarkable properties.

Joint work with Coralia Cartis (University of Oxford) and Nick Gould (STFC-RAL).

Miroslav Tůma *Algebraic preconditioning of symmetric indefinite systems*

Sparse symmetric indefinite linear systems of equations arise in many practical applications. A preconditioned iterative method is frequently the method of choice. In the talk we will consider both general indefinite systems and saddle-point problems and we will summarize our work along these lines based on the recently adopted limited memory approach. A number of new ideas proposed with the goal of improving the stability, robustness and efficiency of the resulting preconditioner will be mentioned.

Based on joint work with Jennifer Scott

Bora Uçar *Two approximation algorithms for bipartite matching on multicore architectures*

We propose two heuristics for the bipartite matching problem that are amenable to shared-memory parallelization. The first heuristic is very intriguing from a parallelization perspective. It has no significant algorithmic synchronization overhead and no conflict resolution is needed across threads. We show that this heuristic has an approximation ratio of around 0.632 under some common conditions. The second heuristic is designed to obtain a larger matching by employing the well-known Karp-Sipser heuristic on a judiciously chosen subgraph of the original graph. We show that the Karp-Sipser heuristic always finds a maximum cardinality matching in the chosen subgraph. Although the Karp-Sipser heuristic is hard to parallelize for general graphs, we exploit the structure of the selected subgraphs to propose a specialized implementation which demonstrates very good scalability. We prove that this second heuristic has an approximation guarantee of around 0.866 under the same conditions as in the first algorithm. We discuss parallel implementations of the proposed heuristics on a multicore architecture. Experimental results, for demonstrating speed-ups and verifying the theoretical results in practice, are provided.

Joint work with Fanny Dufosse, Inria Lille, Nord Europe and Kamer Kaya, Sabanci University, Istanbul, Turkey

Pascal Vezolle *OpenPower and IBM Exascale hybrid architecture overview and benefits for sparse matrices*

Over the last several decades, the computing power of processors has roughly doubled every two years. Initially, increasing clock frequency was the main enabler of increased performance. When the CPU clock frequency flattened, the introduction of multiple processing cores became the major driver. This approach has increased significantly the computing capacities versus the memory performance per core. The accelerator technology massive adoption in HPC, like Graphic Processor Unit (GPU), has reinforced the computation and memory imbalance. This imbalance is a major performance bottleneck for most of the applications based of large sparse matrices. In this presentation we will outline IBM coming HPC pre-exaflop innovative architecture involving multiple heterogeneous processing elements, different memory pools with non-uniform memory access (NUMA). We will show benefits for sparse matrix implementations of high memory bandwidth, and CPU/GPU NUMA integration with no data programmable transfer. We will illustrate through a simple sparse matrix-vector product example the memory dependency and implicitly the performance degradation depending on the number of cores used per node.

Luis Nunes Vicente *Direct search based on deterministic and probabilistic descent*

Direct-search methods are a class of popular derivative-free algorithms characterized by evaluating the objective function using a step size and a number of (polling) directions. When applied to the minimization of smooth functions, the polling directions are typically taken from positive spanning sets which in turn must have at least $n+1$ vectors in an n-dimensional variable space. In addition, to ensure the global convergence of these algorithms, the positive spanning sets used throughout the iterations must be uniformly non-degenerate in the sense of having a positive (cosine) measure bounded away from zero.

However, recent numerical results indicated that randomly generating the polling directions without imposing the positive spanning property can improve the performance of these methods, especially when the number of directions is chosen considerably less than $n + 1$.

In this talk, we analyze direct-search algorithms when the polling directions are probabilistic descent, meaning that with a certain probability at least one of them is of descent type. Such a framework enjoys almost-sure global convergence. More interestingly, we will show a global decaying rate of $1/\sqrt{k}$ for the gradient size, with overwhelmingly high probability, matching the corresponding rate for the deterministic versions of the gradient method or of direct search. Our analysis helps to understand numerical behavior and the choice of the number of polling directions.

We will also review global rates and worst case complexity bounds for derivative-free optimization using direct search as a guide.

This is joint work with Mahdi Dodangeh, R. Garmanjani, S. Gratton, Clément Royer, and Zaikun Zhang.

Jianlin Xia *Superfast direct eigensolvers for large symmetric matrices and the structured perturbation analysis.*

We consider the eigenvalue solution of some large symmetric matrices, including banded matrices and matrices with small off-diagonal numerical ranks. We present a superfast (nearly $O(n)$ complexity) divide-and-conquer algorithm for finding all the eigenvalues as well as all the eigenvectors (in a structured form), together with the approximation error analysis due to the off-diagonal compression. The matrices can be represented or approximated by certain hierarchical structured forms. We show how to preserve the hierarchical structure throughout the dividing process after recursive updates, and how to quickly perform stable eigendecompositions of the structured forms. The rank structure of the eigenmatrix is also shown. We further discuss the structured perturbation analysis, i.e., how the compression of the off-diagonal blocks impacts the accuracy of the eigenvalues. General results on the approximation errors as well as the accuracy refinement are presented. They show that structured methods can serve as an effective and efficient tool for approximate eigenvalue solutions with controllable accuracy. In particular, some eigenvalues may be accurately estimated even though the off-diagonal approximation accuracy is modest.

The algorithm and analysis are useful for finding the eigendecomposition of matrices arising from some important applications, such as banded matrices, Toeplitz matrices, and certain sparse or dense discretized problems. Extensions to SVDs and repeated eigenvalues are considered. The algorithm can also be made matrix-free, i.e., the structured or sparse approximations can be constructed based on matrix-vector multiplications only.

Joint work with James Vogel.

Yin Zhang *Block algorithms in an augmented Rayleigh-Ritz framework for large-scale exterior eigenpair computation.*

Iterative algorithms for computing a set of eigenpairs of large matrices consist of two main steps: a subspace update step and a Rayleigh-Ritz (RR) projection step. In this paper, we propose an augmented Rayleigh-Ritz (ARR) step that can provably accelerate convergence under mild conditions. We consider two block (as opposed to Krylov subspace) algorithms by coupling the ARR procedure with two subspace update schemes: (i) the classic power method applied to multiple vectors without periodic orthogonalization, and (ii) a recently proposed Gauss-Newton method. In block algorithms, the RR step is arguably the bottleneck in scalability as the number of computed eigenpairs increase. Our key design objective is for the algorithms to approach a certain optimal scalability under favorable conditions. That is, they should ideally call the augmented Rayleigh-Ritz step once and attain a sufficient accuracy, while the subspace update step is close to being embarrassingly parallel under a suitable data mapping scheme. We perform extensive computational experiments in Matlab (without explicit code parallelization) to evaluate the proposed algorithms in comparison to a few state-of-the-art eigensolvers. Numerical results suggest strong potentials for the proposed algorithms to reach high levels of scalability on a wide range of problems.

Joint work with Zaiwen Wen.

Posters

Loïc Michel *Derivative-free optimization by model-free control approach.*

The model-free control approach has been designed as a robust controller to overcome difficulties of tuning classical controller when the process to control is not "well-modeled". The proposed controller is very easy to tune considering a few parameters that are related to a rough estimation of the general behavior of the process to control (like dc gain...). Therefore, no specific identification of the process is required while providing very good tracking performances for nonlinear systems. To ensure such interesting performances, the controller needs the computation of numerical derivatives of the measured controlled output signal, which could be difficult to obtain especially in a noisy environment. Recent progresses substituted the computation of the numerical derivatives by an initialization function that makes the controller more robust when stabilizing for example switched processes and such. Last developments would extend the properties of the proposed model-free controller to include the extremum seeking control of nonlinear systems without any computation of derivative or gradient. The proposed derivative-free optimization strategy has been tested in the case of simple convex and nonconvex systems that may include linear constraints. Moreover, the tracking of the extremum is also possible when changes in the nonlinear system occur in real-time.

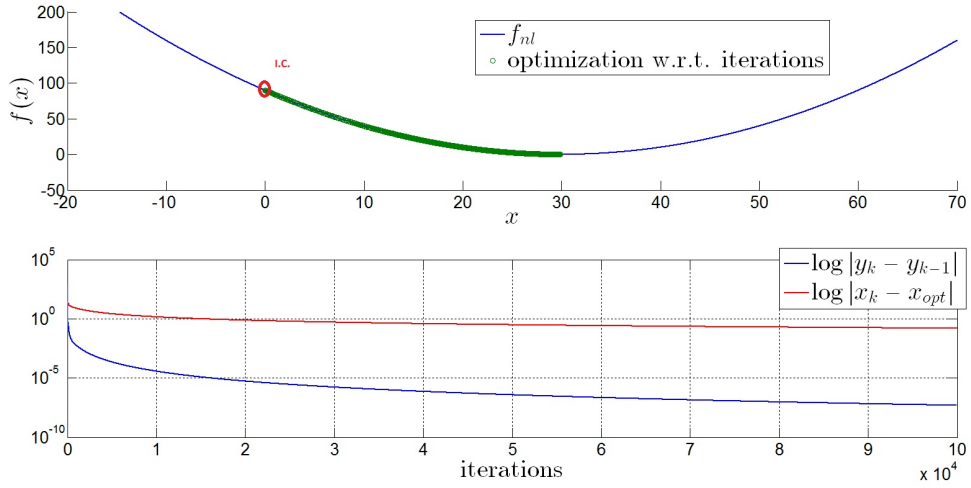

Figure 1: Example of minimum tracking of a convex function.

David Titley-Peloquin *The componentwise symmetric backward error for sparse symmetric systems of equations.*

At the Householder Symposium XIX on Numerical Linear Algebra held last summer in Spa, Belgium, an 'open problems' session was organized. During this session, James Demmel presented the following challenge: solving the *symmetric* componentwise relative backward error problem for linear systems of equations. We decided to take on this challenge. In this poster we present some progress we have made towards its solution.

Joint work with Serge Gratton (CERFACS-IRIT joint Lab)

Sponsored also by