Sparse Days

September 6-7th, 2011

CERFACS, Toulouse, France

Tuesday, September 6th 2011

9.30 - 10.00 Registration and coffee break

Session I

10.00 - 10.30 *An introductory look at the antibandwidth problem*
J. A. Scott (Rutherford Appleton Laboratory (RAL), UK)

10.30 - 11.00 *Balancing to prescribed row and column sums*
P. Knight (Department of Mathematics and Statistics, University of Strathclyde Glasgow, Scotland)

11.00 - 11.05 Break

11.05 - 11.35 *A combinatorial problem in sparse orthogonal factorization*
E. G. Ng (Lawrence Berkeley National Laboratory, USA)

11.35 - 12.05 *On the memory behaviour of a multifrontal QR software for multicore systems*
A. Buttari (CNRS-IRIT, Toulouse, France)

Lunch

Session II

14.00 - 14.30 *On direct elimination of constraints in KKT systems*
C. Ashcraft (Livermore Software Technology Corporation, USA)

14.30 - 15.00 *A scalable Helmholtz solver combining the deflation with shifted Laplace preconditioner*
A. H. Sheikh (Delft University of Technology, Netherlands)

15.00 - 15.30 *Sparse matrix computations in arterial fluid mechanics*
M. Manguoglu (Department of Computer Engineering, Middle East Technical University Ankara, Turkey)

15.30 - 16.00 Coffee break

Session III

16.00 - 16.30 *CALU-PRRP: a communication avoiding LU factorization algorithm with Panel Rank Revealing Pivoting*
A. Khabou (Laboratoire de Recherche en Informatique, Université Paris-Sud XI, INRIA Saclay - Ile de France, France)

16.30 - 17.00 *The Build to Order Compiler for Automating Matrix Algebra–Moving into Sparse Matrices*
E. R. Jessup (Department of Computer Science, University of Colorado at Boulder, USA)

19.30 Banquet at the restaurant Le Pôvre Yves in Toulouse.

Wednesday, September 7th 2011

Session IV

9.30 - 10.00 *Towards a scalable parallel sparse linear system solver*
A. H. Sameh (Department of Computer Science, Purdue University, USA)

10.00 - 10.30 *On algorithms for the maximum cardinality matching problem in bipartite graphs*
B. Uçar (ENS-Lyon, France)

10.30 - 11.00 Coffee break

Session V

11.00 - 11.30 *Memory affinity in sparse matrix-vector multiplications on multi-core architectures*
M. Sosonkina (Ames Laboratory and Iowa State University, USA)

11.30 - 12.00 *Parallel multigrid solver for time harmonic Maxwell equations*
M. Chanaud (INRIA Bordeaux Sud-Ouest, France)

Lunch

Session VI

13.30 - 14.00 *Backward error and its estimates in linear least squares problems*
P. Jiranek (IRIT, Toulouse, France)

14.00 - 14.30 *BFD: a preconditioning technique using block filtering decomposition*
R. Fezzani (INRIA Saclay - Ile de France, France)

Closure

September 6th - Session I

An introductory look at the antibandwidth problem

J. Scott (Rutherford Appleton Laboratory (RAL), UK). The antibandwidth maximization problem is the dual of the well-known bandwidth minimization problem. In this talk, we introduce the problem and consider the feasibility of adapting heuristic algorithms for the bandwidth minimization problem to the antibandwidth maximization problem. In particular, using an inexpensive level-based heuristic we obtain an initial ordering that we refine using a hill-climbing algorithm. This approach performs well on matrices coming from a range of practical problems with underlying grids. Comparisons with existing approaches show that, on this class of problems, our algorithm can be competitive with recently reported results in terms of quality while being significantly faster.

Balancing to prescribed row and column sums

P. A. Knight (Department of Mathematics and Statistics, University of Strathclyde, Glasgow, Scotland) joint work with D. Ruiz (INPT-ENSEEIHT, Toulouse, France). We consider the following problem: given $A \in \mathbb{R}^{m \times n}$, $p \in \mathbb{R}^m$, $q \in \mathbb{R}^n$, $A, p, q \geq 0$, find diagonal matrices D_1 and D_2 so that $P = D_1 A D_2$ satisfies the equalities $Pe = p$, $P^T e = q$. That is, we attempt to balance a matrix so that it has prescribed row and column sums. Applications where we might want P include the interpretation of economic data, preconditioning, understanding traffic circulation, assigning seats fairly after elections, matching protein samples and ordering nodes in a graph. In many of these applications, P will be large and sparse.

We investigate the most well-known method for solving this problem (a variant of the Sinkhorn-Knopp algorithm) proving a new convergence result, and investigate a method which is potentially much faster, based on an inexact Newton method incorporating a preconditioned conjugate gradient iteration.

This method almost invariably performs better than the Sinkhorn-Knopp approach, and if A is symmetric the performance is impressive. However, if A is unsymmetric, the conjugate gradient iteration can stagnate. In part this is a consequence of the fact that the linear systems we solve are singular (albeit consistent). We analyse the causes in detail and describe our attempts to overcome them.

A combinatorial problem in sparse orthogonal factorization

E. G. Ng (Lawrence Berkeley National Laboratory, USA). Let A be an m by n sparse matrix, with $m \geq n$. We consider the problem of finding a k by n submatrix B of A, $k \leq m$, so that the orthogonal factorization of B is sparser than that of A. The orthogonal factorization of B can be useful in the solution of the overdertermined linear system $Ax = b$ using the least squares method. In this talk, we discuss several heuristics for finding B.

On the memory behaviour of a multifrontal QR software for multicore systems

A. Buttari (CNRS-IRIT, Toulouse, France). The advent of multicore processors represents a disruptive event in the history of computer science as conventional parallel programming paradigms are proving incapable of fully exploiting their potential for concurrent computations. The need for different or new programming models clearly arises from recent studies which identify fine-granularity and dynamic execution as the keys to achieve high efficiency on multicore systems. These models can be effectively applied to the multifrontal method for the QR factorization of sparse matrices providing a very high efficiency achieved through a fine-grained partitioning of data and a dynamic scheduling of computational tasks relying on a dataflow parallel programming model. In this context, we will present a technique for the scheduling of computational tasks whose objective is to maximize the locality of data in NUMA multicore systems. We will present experimental result showing how the multifrontal QR factorization benefits from the resulting reduction in data transfers. Finally we will discuss possible improvements to the scheduling technique which are currently under investigation.

September 6th - Session II

On direct elimination of constraints in KKT systems

C. Ashcraft (Livermore Software Technology Corporation, USA). Many applications in engineering and scientific computing require "solving" a linear system, where we want the solution u to satisfy some constraints $Cu = f$ exactly as well as mininize the norm of the residual of the linear system $\|f - Ku\|_2$.

Engineers typically have used *direct elimination* to solve this coupled system of linear equations. Mathematicians are more inclined to view this technique as a null space projection method. The two are equivalent.

We form a KKT system using the stiffness and constraint linear systems. When the stiffness matrix is positive semidefinite and the constraint matrix has full rank, the KKT matrix is indefinite. Using direct elimination (or null space projection), we convert the $(n + r) \times (n + r)$ indefinite KKT matrix to a $(n - r) \times (n - r)$ positive definite matrix.

The method can be viewed as a pre-ordering and a partial factorization of a KKT system formed of the stiffness and constraint linear systems. This approach requires analyzing the constraint matrix to find row and column permutations. This is known as the "nice basis problem".

We compute the inverse of a sparse nonsingular submatrix of the constraints, and use this to compute a modified but equivalent constraint system. We then form a simpler KKT system, and eliminate the constraint rows with an equal number of rows of the stiffness matrix to arrive at the reduced positive definite linear system.

We will discuss these issues within the context of a large commerical finite element analysis code. Our typical constraint linear systems have special structure of which we can take advantage in the analysis phase. The computation of the reduced linear system in distributed memory mode is also challenging.

A scalable Helmholtz solver combining the deflation with shifted Laplace preconditioner

A. H. Sheikh (Delft University of Technology, Netherlands) joint work with D. Lahaye (Delft University of Technology, Netherlands) and K. Vuik (Delft University of Technology, Netherlands). Our object is to develop high performance iterative solution algorithm for solving the discrete Helmholtz equation modeling wave propagation on large scale. Ingredients in our work are the shifted Laplace preconditioner and deflation. The development of the shifted Laplace preconditioner for the Helmholtz equation was a breakthrough in the development of efficient solution techniques for the Helmholtz equation. The distinct feature of this preconditioner is the introduction of a complex shift, effective introducing damping of wave propagation in the approximate solve. This preconditioner was extensively discussed in various texts and applied in a number of different contexts. Although performant, the resulting algorithm is not truly scalable. The bigger the wavenumber, the more spectrum scatters away from one, hampering the convergence. Idea of projection has been used since long to deflate unfavorable eigenvalues. By inducting eigenvectors corresponding to unwanted eigenvalues, better convergence for CG and GMRES has been reported in various texts. We also combine the idea of deflation with shifted Laplace preconditioner, which leads to a scalable Helmholtz solver, in the sense iterations does not depend upon parameters. We provide a convergence analysis. We perform a Fourier two-grid analysis of one-dimensional model problem with Dirichlet boundary conditions discretized by a second order accurate finite difference scheme. The components analyzed are the shifted Laplace preconditioner used as smoother, full-weighting and linear interpolation intergrid transfer operators, and a Galerkin coarsening scheme. This Fourier analysis results in a closed form expression for the eigenvalues of the two-grid operator. This expressions shows that the spectrum is favourable for convergence of Krylov subspace methods. We apply the deflated shifted Laplace preconditioner to two-dimensional model problems method with constant and non-constant wave numbers and Sommerfeld boundary conditions discretized by second order accurate finite difference scheme on uniform meshes. Numerical results show that the number of GMRES iterations is wave-number independent.

Sparse matrix computations in arterial fluid mechanics

M. Manguoglu (Department of Computer Engineering, Middle East Technical University, Ankara, Turkey) joint work with K. Takizawa (Department of Modern Mechanical Engineering and Waseda Institute for Advanced Study, Waseda University, 1-6-1 Nishi-Waseda, Shinjuku-ku,Tokyo 169-8050, Japan), A. H. Sameh (Department of Computer Science, Purdue University, West Lafayette, Indiana, USA) and T. Tezduyar (Mechanical Engineering, Rice University MS 321, 6100 Main Street, Houston, TX 77005, USA).

Iterative solution of large sparse nonsymmetric linear equation systems is one of the numerical challenges in arterial fluidstructure interaction computations. This is because the fluid mechanics parts of the fluid-structure block of the equation system that needs to be solved at every nonlinear iteration of each time step corresponds to incompressible flow, the computational domains include slender parts, and accurate wall shear stress calculations require boundary layer mesh refinement near the arterial walls. We propose a hybrid parallel sparse algorithm, Domain-Decomposing Parallel Solver - DDPS [1, 2, 3] to address this challenge. As the test case, we use a fluid mechanics equation system generated by starting with an arterial shape and flow field coming from an FSI computation and performing two time steps of fluid mechanics computation with a prescribed arterial shape change, also coming from the FSI computation. We show how the DDPS algorithm performs in solving the equation system and demonstrate the scalability of the algorithm.

References

[1] M. Manguoglu and K. Takizawa and A. Sameh and T. Tezduyar, *A parallel sparse algorithm targeting arterial fluid mechanics computations*, Computational Mechanics, pp. 18, 2011.

[2] M. Manguoglu and K. Takizawa and A. Sameh and T. Tezduyar, *Nested and parallel sparse algorithms for arterial fluid mechanics computations with boundary layer mesh refinement*, International Journal for Numerical Methods in Fluids, vol. 65, pp. 135149, 2011.

[3] M. Manguoglu, *A domain decomposing parallel sparse linear system solver*, Journal of Computational and Applied Mathematics, in review.

September 6th - Session III

CALU-PRRP: a communication avoiding LU factorization algorithm with Panel Rank Revealing Pivoting

A. Khabou (Laboratoire de Recherche en Informatique Université Paris-Sud 11, INRIA Saclay - Ile de France) joint work with J. W. Demmel (Computer Science Division and Mathematics Department, UC Berkeley, USA), L. Grigori (INRIA Saclay - Ile de France, Laboratoire de Recherche en Informatique Université Paris-Sud 11, France) and M. Gu (Mathematics Department, UC Berkeley, USA). We present the LU decomposition with panel rank revealing pivoting (LU_PRRP), an LU factorization algorithm based on a new pivoting strategy that performs Strong Rank Revealing QR on the panels to choose the pivot rows. LU_PRRP is more stable than Gaussian elimination with partial pivoting (GEPP), with a theoretical pivot growth upper bound of $(1+2b)^{\frac{n}{b}}$, where b is the size of the panel used during the factorization. For example, if the size of the panel is $b = 64$, then $(1 + 2b)^{n/b} = (1.079)^n \ll 2^{n-1}$, where 2^{n-1} is the pivot growth upper bound for GEPP. Our extensive numerical experiments show that the new pivoting scheme is as numerically stable as GEPP but is more resistant to pathological cases and easily beats the Kahan matrix. The LU_PRRP method only does $O(n^2)$ flops more than GEPP.

We also present CALU_PRRP, a communication avoiding version of LU_PRRP that is optimal in terms of communication. Like the CALU algorithm, this algorithm is based on tournament pivoting, but is more stable than CALU in terms of worst case pivot growth.

The Build to Order Compiler for Automating Matrix Algebra–Moving into Sparse Matrices

E. R. Jessup (Department of Computer Science, University of Colorado at Boulder, USA) joint work with I. Karlin, T. Nelson, and P. Zelinsky (Department of Computer Science, University of Colorado at Boulder, USA), J. Siek and G. Belter (Department of Electrical, Computer, and Energy Engineering, University of Colorado at Boulder, USA) and B. Norris (Mathematics and Computer Science Division, Argonne National Laboratory, USA) Data movement limits the performance of the matrix algebra calculations used in many scientific programs. We have developed the Build to Order (BTO) compiler to automate tuning that reduces memory traffic in those computations. The optimizations included are loop fusion and cache blocking and data partitioning to enable the creation of shared memory parallel codes. Within BTO, an analytic memory model efficiently and accurately reduces the number of serial loop fusion options considered. The result is efficient linear algebra kernels that run over 100% faster than vendor optimized BLAS on serial and parallel machines. The initial version of BTO operates very successfully on dense matrice. Preliminary results suggest that performance gains are not as great for sparse matrices as they are for dense, but fully understanding the issues surrounding sparse matrices is a matter of future work. We invite comment from and collaboration with practitioners of sparse matrix algebra.

September 7th - Session IV

Towards a scalable parallel sparse linear system solver

A. H. Sameh (Department of Computer Science, Purdue University, West Lafayette, Indiana, USA) joint work with M. Manguoglu (Department of Computer Engineering, Middle East Technical University, Ankara, Turkey), O. Schenk (Department of Computer Science, University of Basel, Switzerland).

Designing sparse linear system solvers capable of achieving high parallel scalability on computing platforms with thousands of cores is a challenging task. In this paper we present a hybrid algorithm that is more scalable than current parallel direct sparse solvers and more robust than approximate LU-factorization-, or algebraic multigrid-, preconditioned Krylov subspace methods. The proposed scheme combines features from the SPIKE family of banded solvers, and the direct sparse solver PARDISO. A vital first step is the creation of a scalable parallel reordering scheme that enables the extraction of an effective banded preconditioner (dense or sparse within the band). This is followed by outer Krylov subspace iterations in which systems involving the preconditioner are handled via a hybrid solver PSPIKE that utilizes a specialized version of the sparse direct solver PARDISO. Numerical experiments that demonstrate the robustness and parallel scalability of this hybrid solver are presented.

References

[1] M. Naumov and M. Manguoglu and A. Sameh, *A tearing-based hybrid parallel sparse linear system Solver*, JCAM, Vol. 234, pp. 3025-3038, 2010.

[2] M. Manguoglu and M. Koyuturk and A. Sameh and A. Grama, *Weighted Matrix Ordering and Parallel Banded Preconditioners for Iterative Linear System Solvers*, SIAM Journal on Scientific Computing, Volume 32, pp.1201-1216, 2010.

[3] M. Manguoglu and A. Sameh and O. Schenk, *PSPIKE: A Parallel Hybrid Sparse Linear System Solver*, Lecture Notes in Computer Science, Volume 5704, pp.797-808, 2009.

[4] O. Schenk and M. Manguoglu and A. Sameh and M. Christen and M. Sathe, *Parallel Scalable PDE-Constrained Optimization: Antenna Identification in Hyperthermia Cancer Treatment Planning*, Computer Science Research and Development, Volume 23, pp. 177-183, 2009.

On algorithms for the maximum cardinality matching problem in bipartite graphs

B. Uçar (ENS-Lyon, France) joint work with I. S. Duff (CERFACS, France) and K. Kaya (CERFACS, France). We investigate exact and heuristic algorithms for the cardinality matching problem in bipartite graphs. On the exact algorithms side we discuss and evaluate seven algorithms based on augmenting path searches, and one based on push-relabel techniques. We identify two of these algorithms as the best on a wide range of real life problems. For the heuristic algorithms side we recall some well known ones and discuss their merits.

September 7th - Session V

Memory affinity in sparse matrix-vector multiplications on multi-core architectures

M. Sosonkina (Ames Laboratory and Iowa State University, USA) joint work with A. Srinivasa (Ames Laboratory and Iowa State University, USA). As the core counts on multi-processor systems increase, so does the memory contention when the cores try to access main memory simultaneously. To circumvent this problem, modern systems are moving increasingly towards Non-Uniform Memory Access (NUMA) architectures, in which the physical memory is split into multiple (typically 2-4) banks, also called nodes. Each node is associated with a processor or a set of cores which has the fastest access to the memory on the local node while maintaining the shared virtual address space. Thus, if the data shared by certain threads is located on the node local to these threads, the remote memory access and bandwidth contention is diminished. A policy of data placement within the NUMA nodes is often called memory affinity.

This talk will demonstrate some memory affinity performance effects on applications that employ multi-threaded sparse matrix-vector multiplication (mt-SpMV). Next, the strategies for judicial shared data placement will be investigated along with their dependence on sparse matrix format and thread scheduling. The experiments were conducted with the Conjugate Gradient (CG) NAS benchmark and with the realistic nuclear structure calculations performed by the MFDn code on the Cray XE6 Hopper supercomputer at the National Energy Research Scientific Computing Center (NERSC). The proposed memory affinity strategies increase the performance gains as the number of threads grows. In the mt-SpMV dominated CG using 12 threads, an almost twofold improvement has been observed. This speed-up grew linearly with the number of threads.

Parallel multigrid solver for time harmonic Maxwell equations

M. Chanaud (INRIA Bordeaux Sud-Ouest, France) joint work with D. Goudin (CEA/DAM/ CESTA, France), J-J. Pesqué (CEA/DAM/CESTA, France), L. Giraud (INRIA Bordeaux Sud-Ouest, France) and J. Roman (INRIA Bordeaux Sud-Ouest, France). Our goal is to develop a high performance parallel solver based on a geometric multigrid method for time harmonic Maxwell equations $\nabla\times\nabla\times E - \kappa^2 E = 0$. These equations are discretized with Nédélec tetrahedral first order finite elements and solved by PaStiX direct solver in actual CEA electromagnetic simulation code. Due to the relation between the wavelength of the incident wave and the mesh mean edge length, the mesh and linear system sizes increase with the frequency, the object's material properties (permeability and permittivity) and the object size. As a consequence, simulations require linear systems with many millions of unknowns. In order to replace the actual direct solver, we develop a geometric multigrid method driven by a direct solver using the full-multigrid scheme : the direct solution of the problem on the initial coarse mesh is interpolated on a fine mesh automatically generated from the coarse mesh. The interpolated vector is the initial guess for the fine mesh solution which is computed by a matrix-free Jacobi iterative solver. Using the low memory consumption and the scalability of the matrix-free Jacobi solver we are able to solve large linear systems with billions of unknowns. This talk will present the multigrid method, convergence results on either high frequency and complex material problems, the convergence of the multigrid-preconditioned GMRES, and demonstrate the solver scalability.

September 7th - Session VI

Backward error and its estimates in linear least squares problems

P. Jiranek (IRIT, Toulouse, France) joint work with D. Titley-Peloquin (Mathematical Institute at the University of Oxford, UK) and S. Gratton (ENSEEIHT-IRIT, Toulouse, France). We are interested in computing the backward error associated with an approximation of the solution of a linear least squares problem, which is given by the minimal singular value of certain large matrix and is hard to evaluate directly. We analyze the accuracy of several estimates of the LS backward error proposed in literature, which can be cheaply computed in practice and discuss their use in iterative methods.

BFD: a preconditioning technique using block filtering decomposition

R. Fezzani (INRIA Saclay - Ile de France). We introduce a new preconditioning technique that is suitable for matrices arising from the discretization of a system of PDEs on unstructured grids. The preconditioner satisfies a so-called filtering property, which ensures that the input matrix is identical with the preconditioner on a given filtering vector. This vector is chosen to alleviate the effect of low frequency modes on convergence and so decrease or eliminate the plateau which is often observed in the convergence of iterative methods. In particular, we present a general approach that allows to ensure that the filtering condition is satisfied in a matrix decomposition. The input matrix can have an arbitrary sparse structure. Hence, it can be reordered using nested dissection, to allow a parallel computation of the preconditioner and of the iterative process.

Last update: September 5, 2011