The Schur Complement Systems in the Mixed Hybrid Finite Element Approximation of the Potential Fluid Flow Problem

Maryška, Jiří
1997

SCHUR COMPLEMENT SYSTEMS IN THE MIXED-HYBRID FINITE ELEMENT APPROXIMATION OF THE POTENTIAL FLUID FLOW PROBLEM *

J. MARYŠKA, M. ROZLOŽNÍK † AND M. TŮMA‡

Abstract. The mixed hybrid finite element discretization of Darcy's law and continuity equation describing the potential fluid flow problem in porous media leads to a symmetric indefinite linear system for the pressure and velocity vector components. As a method of solution the reduction to three Schur complement systems based on successive block elimination is considered. The first and second Schur complement matrices are formed eliminating the velocity and pressure variables, respectively and the third Schur complement matrix is obtained by elimination of a part of Lagrange multipliers that come from the hybridization of a mixed method. The structural properties of these consecutive Schur complement matrices in terms of the discretization parameters are studied in detail. Based on these results the computational complexity of a direct solution method is estimated and compared to the computational cost of the iterative conjugate gradient method applied to Schur complement systems. It is shown that due to special block structure the spectral properties of successive Schur complement matrices do not deteriorate and the approach based on the block elimination and subsequent iterative solution is well justified. Theoretical results are illustrated by numerical experiments.

1. Introduction. Let Ω be a bounded domain in $\mathcal{R}^3$ with a Lipschitz continuous boundary $\partial\Omega$. The potential fluid flow in saturated porous media can be described by the velocity $\mathbf{u}$ using Darcy's law and by the continuity equation for incompressible flow

$$\mathbf{A}\mathbf{u} = -\nabla p, \quad \nabla \cdot \mathbf{u} = q, \tag{1.1}$$

where p is the piezometric potential (fluid pressure), $\mathbf{A}$ is a symmetric and uniformly positive definite second rank tensor of the hydraulic resistance of medium with $[\mathbf{A}(\mathbf{x})]_{ij} \in L^\infty(\Omega)$ for all $i, j \in \{1, 2, 3\}$ and q represents the density of potential sources in the medium. The boundary conditions are given by

$$p = p_D \quad \text{on} \quad \partial\Omega_D, \quad \mathbf{u} \cdot \mathbf{n} = u_N \quad \text{on} \quad \partial\Omega_N, \tag{1.2}$$

where $\partial\Omega = \overline{\partial\Omega_D} \cup \overline{\partial\Omega_N}$ are such that $\partial\Omega_D \neq \emptyset$, $\partial\Omega_D \cap \partial\Omega_N = \emptyset$ and $\mathbf{n}$ is the outward normal vector defined (almost everywhere) on the boundary $\partial\Omega$.

Assume that the domain Ω is a polyhedron and it is divided into a collection of subdomains such that every subdomain is a trilateral prism with vertical faces and general nonparallel bases (see, e.g., [11], [14] or [15]). We will denote the discretization of the domain Ω by $\mathcal{E}_h$ and assume an uniform regular mesh with the discretization parameter h. Denote also the collection of all faces of elements $e \in \mathcal{E}_h$ which are not adjacent

* This work was supported by the Grant Agency of the Czech Republic under grant 201/98/P108 and by the grant AS CR A2030706. Revised version October 1999.

† Seminar for Applied Mathematics, Swiss Federal Institute of Technology (ETH) Zurich, ETH-Zentrum, CH-8092 Zurich, Switzerland. (miro@sam.math.ethz.ch)

‡ Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic. (maryska@uivt.cas.cz, miro@uivt.cas.cz, tuma@uivt.cas.cz)

to the boundary $\partial\Omega_D$ by $\Gamma_h = \cup_{e\in\mathcal{E}_h}\partial e - \partial\Omega_D$ and introduce the set of interior faces $int(\Gamma_h) = \Gamma_h - \partial\Omega_N$.

We consider the following low order finite element approximation. Let

$$\mathbf{RT}^0(e) = \{\mathbf{v}^e;\ \mathbf{v}^e(\mathbf{x}) = \sum_{j=1}^{5}\nu_j\mathbf{v}_j^e(\mathbf{x}),\ \forall\mathbf{x} = (x_1,x_2,x_3)\in e\}, \tag{1.3}$$

be the space spanned by the linearly independent basis functions $\mathbf{v}_j^e(\mathbf{x}),\ j = 1,...,5$, defined on the element $e\in\mathcal{E}_h$ in the form

$$\mathbf{v}_j^e(\mathbf{x}) = k_j^e\begin{pmatrix}0\\0\\x_3-\alpha_{j3}^e\end{pmatrix},\ j=1,2,\quad \mathbf{v}_j^e(\mathbf{x}) = k_j^e\begin{pmatrix}x_1-\alpha_{j1}^e\\x_2-\alpha_{j2}^e\\\beta_j^e x_3-\alpha_{j3}^e\end{pmatrix},\ j=3,4,5 \tag{1.4}$$

and such that they are orthonormal with respect to the set of functionals

$$\mathcal{F}_k(\mathbf{v}_j^e) = \int_{f_k^e}\mathbf{n}_k^e\cdot\mathbf{v}_j^e dS = \delta_{jk},\ j,k = 1,...,5. \tag{1.5}$$

Here f_k^e denotes the k-th face of the element $e\in\mathcal{E}_h$ and $\mathbf{n}_k^e = (n_{k,1}^e, n_{k,2}^e, n_{k,3}^e)$ is the outward normal vector with respect to the face f_k^e. The velocity function $\mathbf{u}$ will be then approximated by vector functions linear on every element $e\in\mathcal{E}_h$ from the Raviart-Thomas space

$$\mathbf{RT}_{-1}^0(\mathcal{E}_h) = \{\mathbf{v}_h\in\mathbf{L}^2(\Omega);\ \mathbf{v}_h|^e\in\mathbf{RT}^0(e),\ \forall e\in\mathcal{E}_h\}, \tag{1.6}$$

where $\mathbf{v}_h|^e$ denotes the restriction of a function $\mathbf{v}_h$ onto the element $e\in\mathcal{E}_h$. Further denote the space of constant functions on each element $e\in\mathcal{E}_h$ by $M^0(e)$ and denote the space of constant functions on each face $f\in\Gamma_h$ by $M^0(f)$. The piezometric potential p will be approximated by the space which consists of elementwise constant functions

$$M_{-1}^0(\mathcal{E}_h) = \{\phi_h\in L^2(\Omega);\ \phi_h|^e\in M^0(e),\ \forall e\in\mathcal{E}_h\}, \tag{1.7}$$

where $\phi_h|^e$ is the restriction of a function ϕ_h onto element $e\in\mathcal{E}_h$. The Lagrange multipliers coming from the hybridization of a method will be approximated by the space of all functions constant on every face from Γ_h

$$M_{-1}^0(\Gamma_h) = \{\mu_h\in L^2(\Omega);\ \Gamma_h\to R;\ \mu_h|^f\in M^0(f),\ \forall f\in\Gamma_h\}. \tag{1.8}$$

Here $\mu_h|^f$ denotes the restriction of a function μ_h onto the face $f\in\Gamma_h$. Analogously we introduce the spaces $M_{-1}^0(\partial\Omega_D)$ and $M_{-1}^0(\partial\Omega_N)$ as the spaces of functions constant on every face from $\cup_{e\in\mathcal{E}_h}\partial e\cap\partial\Omega_D$ and $\Gamma_h\cap\partial\Omega_N$, respectively. The detailed description of the spaces that we use can be found in [14] (see also [11] or [15]).

The Raviart-Thomas approximation of the mixed-hybrid formulation for the problem (1.1) and (1.2) reads as follows (see [4]):

$$\text{Find}\ \ (\mathbf{u}_h,p_h,\lambda_h)\in\mathbf{RT}_{-1}^0(\mathcal{E}_h)\times M_{-1}^0(\mathcal{E}_h)\times M_{-1}^0(\Gamma_h)\ \text{such that}$$

$$\sum_{e\in\mathcal{E}_h}\{(\mathbf{A}\mathbf{u}_h,\mathbf{v}_h)_{0,e} - (p_h,\nabla\cdot\mathbf{v}_h)_{0,e} + <\lambda_h,\mathbf{n}^e\cdot\mathbf{v}_h>_{\partial e\cap\Gamma_h}\} =$$

$$= <p_{D,h},\mathbf{n}^e\cdot\mathbf{v}_h>_{\partial e\cap\partial\Omega_D};\quad \forall\mathbf{v}_h\in\mathbf{RT}_{-1}^0(\mathcal{E}_h). \tag{1.9}$$

$$(1.10) \qquad -\sum_{e\in\mathcal{E}_h}(\nabla\cdot\mathbf{u}_h,\phi_h)_{0,e}=-(q_h,\phi_h)_{0,\Omega};\quad \forall\phi_h\in M^0_{-1}(\mathcal{E}_h).$$

$$(1.11) \qquad \sum_{e\in\mathcal{E}_h}<\mathbf{n}^e\cdot\mathbf{u}_h,\mu_h>_{\partial e}=<u_{N,h},\mu_h>_{\partial e\cap\partial\Omega_N};\quad \forall\mu_h\in M^0_{-1}(\Gamma_h),$$

where $p_{D,h}$ and $u_{N,h}$ are approximations to the functions p_D and u_N on the spaces $M^0_{-1}(\partial\Omega_D)$ and $M^0_{-1}(\partial\Omega_N)$, respectively; and where the function q is approximated by $q_h\in M^0_{-1}(\mathcal{E}_h)$. For other details we refer to [14] or [11].

Further denote by $NE=|\mathcal{E}_h|$ the number of elements, by $NIF=|int(\Gamma_h)|$ the number of interior inter-element faces and the number of faces with the prescribed Neumann boundary conditions in the discretization by $NNC=|\partial\Omega_N|$. Let $e_i\in\mathcal{E}_h$, $i=1,\dots,NE$, be some numbered ordering of the set of prismatic elements and f_k, $k=1,\dots,NIF+NNC$, be the ordering of the set of non-Dirichlet faces from Γ_h. For every element $e_i\in\mathcal{E}_h$ we denote by NIF_{e_i} the number of interior inter-element faces and by NNC_{e_i} the number of faces with Neumann boundary conditions imposed on the element e_i. Let the finite-dimensional space $\mathbf{RT}^0_{-1}(\mathcal{E}_h)$ be spanned by $NA=5\times NE$ linearly independent basis functions $\mathbf{v}_j$, $j=1,\dots,NA$ from the definition (1.6); let the space $M^0_{-1}(\mathcal{E}_h)$ be spanned by NE linearly independent basis functions ϕ_i, $i=1,\dots,NE$ and finally the space $M^0_{-1}(\Gamma_h)$ be spanned by $NIF+NNC$ linearly independent basis functions μ_k, $k=1,\dots,NIF+NNC$. From this Raviart-Thomas approximation we obtain the system of linear algebraic equations in the form

$$(1.12) \qquad \begin{pmatrix} A & B & C\\ B^T & & \\ C^T & & \end{pmatrix}\begin{pmatrix} u\\ p\\ \lambda\end{pmatrix}=\begin{pmatrix} q_1\\ q_2\\ q_3\end{pmatrix},$$

where $u=(u_1,...,u_{NA})^T$, $p=(p_1,...,p_{NE})^T$, $\lambda=(\lambda_1,...,\lambda_{NIF+NNC})^T$ are unknowns, the symmetric positive definite matrix block $A\in\mathcal{R}^{NA,NA}$ is given by the terms $(\mathbf{A}\mathbf{v}_i,\mathbf{v}_j)_{0,\Omega}$, the outdiagonal block $B\in\mathcal{R}^{NA,NE}$ by $-(\nabla\cdot\mathbf{v}_i,1)_{0,e_j}$ and the block $C\in\mathcal{R}^{NA,NIF+NNC}$ by $<\mathbf{n}_k\cdot\mathbf{v}_i,1>_{f_k}$. Here $\mathbf{n}_k$ is the outward normal vector to with respect to the face $f_k\in\Gamma_h$ (see [11] and [14]).

Let us denote the system matrix in (1.12) by A. The symmetric matrix A is indefinite due to the zero diagonal block of dimension $NBC=NE+NIF+NNC$. The structure of nonzero elements in the matrix from a small model problem can be seen in Figure 1. Partition the submatrix C in A as $(C_1\ C_2)$, where $C_1\in\mathcal{R}^{NA,NIF}$ corresponds to the interior inter-element faces in the discretized domain and $C_2\in\mathcal{R}^{NA,NNC}$ is the face-condition incidence matrix corresponding to the element faces with Neumann boundary conditions. Note that every column of C_1 contains only two nonzero entries equal to 1. The singular values of C_1 are all equal to $\sqrt{2}$ and the matrix block C_2 has orthogonal columns. Moreover, the whole matrix block C has also singular values equal to $\sqrt{2}$ or 1. The matrix B has a special structure. The nonzero elements correspond to the face-element incidence matrix with values equal to -1. Thus all singular values of the matrix B are equal to $\sqrt{5}$ (the matrix B is, up to the normalization coefficients, orthogonal).

It is easy to see from the definition of approximation spaces (see [14] or [15]) that the symmetric positive definite block is 5×5 block diagonal and it was shown in [15]

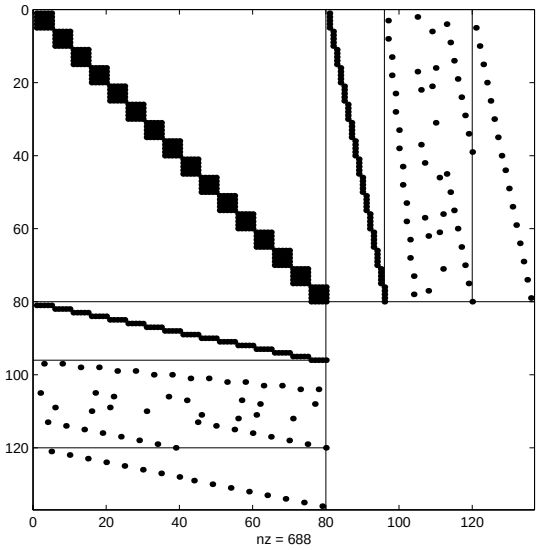

FIG. 1. *Structural pattern of the matrix obtained from mixed hybrid finite element approximation of a model problem with* $h = 1/2$ *(to be discussed in Section 5).*

that the spectrum of the matrix block A satisfies

$$\sigma(A) \subset [c_1 \sqrt[3]{NE}, c_2 \sqrt[3]{NE}], \tag{1.13}$$

where c_1 and c_2 are positive constants independent of the discretization parameters and dependent on the domain and the tensor $\mathbf{A}$. It is also easy to see that the system matrix A in (1.12) is non-singular if and only if the block $(B\ C)$ has a full column rank. Clearly, if the condition $\partial\Omega_D = \emptyset$ holds (all boundary conditions are Neumann conditions), then the matrix block $(B\ C)$ is singular, due to the fact that all sums of row elements are zero. In other words, the function p is unique up to a constant function in the case $\partial\Omega_D = \emptyset$. Assuming $\partial\Omega_D \neq \emptyset$ it follows from the analysis presented in [15] that there exist positive constants c_3 and c_4 such that for the singular values of the matrix block $(B\ C)$ we have

$$sv(B\ C) \subset [c_3/\sqrt[3]{NE}, c_4]. \tag{1.14}$$

Moreover, for eigenvalues of the whole symmetric indefinite matrix A it follows asymptotically $(h \to 0)$

$$\sigma(\mathsf{A}) \subset [-c_5/\sqrt[3]{NE}, -c_6/NE] \cup [c_7 \sqrt[3]{NE}, c_8 \sqrt[3]{NE}], \tag{1.15}$$

where c_5, c_6, c_7 and c_8 are positive constants independent of system parameters.

In this paper, for solving the symmetric indefinite systems (1.12), the successive reduction to Schur complement systems is proposed. We consider three successive Schur complement systems arising during the block elimination of unknowns which correspond to matrix blocks A, B and C_2 respectively, or in other words, which correspond to the elimination of the velocity variables u, the pressure variables p and of a part of the Lagrange multipliers λ. While the concept of reduction to the first and second Schur complement systems is well known as a static condensation (described e.g. in [4], Section V. or in [11])), the proposed reduction to the third Schur complement system seems to

be new. The main contribution of the paper consists in a detailed investigation of the structure of nonzero entries and the spectral properties of the Schur complement matrices. This enables thorough complexity analysis of the direct or iterative solution of corresponding Schur complement systems. A brief analysis of the structure of the first Schur complement matrix can be found in [4] as well as a straightforward observation that its principal leading block is a diagonal. Here we extend this analysis and discuss the mutual relation between the number of nonzero entries in the first Schur complement matrix and the number of nonzeros in the system matrix (1.12). We show further that no fill-in occurs during the process of reduction to the second and third Schur complement system. Moreover, we prove that the number of nonzeros in both these two Schur complements is always less than the number of nonzeros in (1.12). It is shown also that the spectral properties of matrices in such Schur complement systems do not deteriorate during the successive elimination. Thus an approach based on the block reduction and subsequent iterative solution is well justified.

The outline of the paper is as follows. In Section 2, we examine the structural pattern of resulting Schur complement matrices and give estimates for their number of nonzero elements in terms of the discretization parameters listed above. Section 3 is devoted to the solution of the whole indefinite system (1.12) via three Schur complement reductions and subsequent direct solution. Using the graph theoretical results we give the asymptotic bound of the computational complexity for the Cholesky decomposition method applied to the third Schur complement system. In Section 4, we concentrate on the spectral properties of the Schur complement system matrices. The theoretical convergence rate of the iterative conjugate gradient-type method in terms of the discretization parameters is estimated. The asymptotic bounds for the computational work of the iterative solution are given. Section 5 contains some numerical experiments illustrating the previously developed theoretical results. Finally, we give some concluding remarks and mention some open questions for our future work.

2. Structural properties of the Schur complement matrices. In this section we take a closer look to the discretized indefinite system and corresponding Schur complements and we extend the brief analysis from [4]. There are several possibilities for the choice of a block ordering in the consecutive elimination. We shall concentrate on the block ordering which seems to be the most natural and efficient from the point of view of solving the final Schur complement system by a direct solver or by a conjugate gradient-type method. The same ordering for the elimination of the first two blocks was used also in [4], p. 178-181 or in [11]. Note that the static condensation is not the only way to form the successive Schur complements. E.g., in [17] the case of the Raviart-Thomas discretization for the closely related nodal methods was studied and reduction to a different second Schur complement system was discussed.

The following simple result gives the number of nonzero elements in the triangular part of the matrix A. By the triangular part of a matrix M we mean its upper (lower) strict triangle + diagonal. We will deal only with the structural nonzero elements here; we do not take into account accidental cancellations and possible initial zero values of the

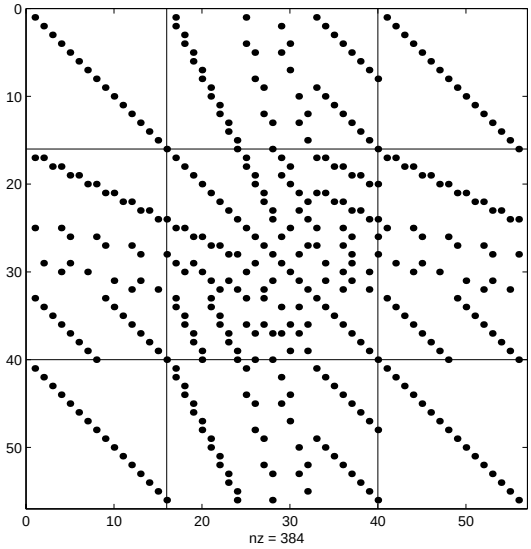

FIG. 2. *Structural pattern of the Schur complement matrix* A/A *for* A *from Figure 1.*

tensor of hydraulic permeability. By the structure of a matrix M we mean $Struct(M) = \{(i,j)|M_{ij} \neq 0\}$.

LEMMA 2.1. *The number of nonzeros in the triangular part of* A *is given by*

$$|sym(A)| = 20NE + 2NIF + NNC. \tag{2.16}$$

Proof. The triangular part of A has $15NE$ nonzeros, the block B contributes by $5NE$ nonzeros, C_1 has $2NIF$ nonzeros and C_2 contains NNC nonzeros. □

The symmetric positive definite matrix block A in (1.12) is block-diagonal, each 5×5 block corresponds to certain element in the discretization of the domain. Therefore it is straightforward to eliminate the velocity variables u and to obtain the first Schur complement system with the matrix

$$\mathsf{A}/A = -\begin{pmatrix} B^T \\ C_1^T \\ C_2^T \end{pmatrix} A^{-1} \begin{pmatrix} B & C_1 & C_2 \end{pmatrix} = -\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & A_{22} & A_{23} \\ A_{13}^T & A_{23}^T & A_{33} \end{pmatrix}.$$

The structure of the matrix A/A for our example problem is shown in Figure 2. For details we also refer to [4], p. 180-181 or [11]. For the number of nonzeros in the matrix A/A we can show the following result.

LEMMA 2.2. *The number of nonzeros in the triangular part of the Schur complement matrix* A/A *is equal to*

$$|sym(\mathsf{A}/A)| = NE + \frac{3}{2}NIF + \frac{3}{2}NNC + \frac{1}{2}\sum_{i \in \mathcal{E}_h}(NIF_i + NNC_i)^2 + \frac{1}{2}\sum_{i \in \mathcal{E}_h}\sum_{j \in Adj(i)} NIF_j.$$

Proof. Clearly, $|sym(A_{11})| = NE$ and $|A_{12}| = |B^T A^{-1} C_1| = 2NIF$. Note that the fill-in for $A^{-1}C_1$ is considerably higher (it is equal to $10NIF$). Further, $|A_{13}| = NNC$

and $|sym(A_{33})| = \frac{1}{2}\sum_{i\in\mathcal{E}_h} NNC_i(NNC_i+1) = \frac{1}{2}NNC + \frac{1}{2}\sum_{i\in\mathcal{E}_h} NNC_i^2$. The number of nonzeros in A_{23} is equal to $\sum_{i\in\mathcal{E}_h} NNC_i\, NIF_i$. Finally, note that

$$Struct(A_{22}) = Struct(C_1^T A^{-1} C_1) = Struct(C_1^T B B^T A^{-1} C_1) = Struct(C_1^T B B^T C_1).$$

Observe that the directed graph of the matrix $B^T C_1$ has the set of arcs

$$E_{B^T C_1} = \{(i,f) \in \mathcal{E}_h \times int(\Gamma_h) |\ f \text{ is an interior face of } i\}.$$

The undirected graph of $C_1^T B B^T C_1$ therefore expresses element-element adjacency relation based on the connectivity through the interior faces inside the domain. It follows that

$$|A_{22}| = \sum_{f\in int(\Gamma_h)} (NIF_{e(f)} + NIF_{\bar{e}(f)} - 1) = \sum_{i\in\mathcal{E}_h} NIF_i(NIF_i - 1) + \sum_{i\in\mathcal{E}_h}\sum_{j\in Adj(i)} NIF_j,$$

where $e(f)$ and $\bar{e}(f)$ are the two elements from $\mathcal{E}_h$ such that $e(f)\cap\bar{e}(f) = \{f\}$. Therefore, considering the relation $2NIF = \sum_{i\in\mathcal{E}_h} NIF_i$ we obtain $|sym(A_{22})| = \frac{1}{2}(|A_{22}| + NIF) = \frac{1}{2}\sum_{i\in\mathcal{E}_h}(NIF_i)^2 + \frac{1}{2}\sum_{i\in\mathcal{E}_h}\sum_{j\in Adj(i)} NIF_j - \frac{1}{2}NIF$. Putting all the partial sums together we get the desired result. □

Consider now the second Schur complement matrix

$$(-\mathsf{A}/A)/A_{11} = -\begin{pmatrix} A_{12}^T \\ A_{13}^T \end{pmatrix} A_{11}^{-1} \begin{pmatrix} A_{12} & A_{13} \end{pmatrix} + \begin{pmatrix} A_{22} & A_{23} \\ A_{23}^T & A_{33} \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{pmatrix}.$$

The structure of $(-\mathsf{A}/A)/A_{11}$ for our example matrix is shown in Figure 3. The matrix block A_{11} in the first Schur complement matrix A/A is diagonal [4], [11]. The following result shows that it is worth to form the Schur complement matrix $(-\mathsf{A}/A)/A_{11}$ from the matrix A/A since no further fill-in appears during the elimination of the block A_{11} corresponding to pressure variables p and so we can further reduce the dimension of the system.

THEOREM 2.1.

$$Struct\left(\begin{pmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{pmatrix}\right) = Struct\left(\begin{pmatrix} A_{22} & A_{23} \\ A_{23}^T & A_{33} \end{pmatrix}\right).$$

Proof. We have the following structural equivalences:

$$\begin{aligned} Struct(B_{11}) &= Struct(A_{22}) + Struct(A_{12}^T A_{11}^{-1} A_{12}) \\ &= Struct(A_{22}) + Struct(C_1^T A^{-1} B A_{11} B^T A^{-1} C_1) \\ &= Struct(A_{22}) + Struct(C_1^T A^{-1} B B^T A^{-1} C_1) \\ &= Struct(A_{22}) + Struct(C_1^T B B^T C_1) = Struct(A_{22}). \end{aligned}$$

$$\begin{aligned} Struct(B_{12}) &= Struct(A_{23}) + Struct(A_{12}^T A_{11}^{-1} A_{13}) \\ &= Struct(A_{23}) + Struct(C_1^T A^{-1} B B^T A^{-1} C_2) \\ &= Struct(A_{23}) + Struct(C_1^T B B^T C_2) \\ &= Struct(A_{23}) + Struct(C_1^T A^{-1} C_2) = Struct(A_{23}). \end{aligned}$$

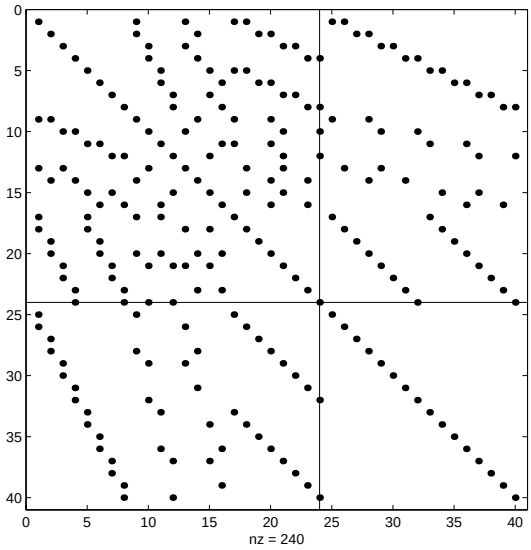

FIG. 3. *Structural pattern of the Schur complement matrix* $(-\mathsf{A}/A)/A_{11}$ *for* A *from Figure 1.*

$$\begin{aligned} Struct(B_{22}) &= Struct(A_{33}) + Struct(A_{13}^T A_{11}^{-1} A_{13}) \\ &= Struct(A_{33}) + Struct(C_2^T A^{-1} B A_{11}^{-1} B^T A^{-1} C_2) \\ &= Struct(A_{33}) + Struct(C_2^T A^{-1} C_2) = Struct(A_{33}). \end{aligned}$$

□

From previous Theorem it is also easy to see that right lower block B_{22} is block-diagonal with blocks of varying size (depending on number of faces with Neumann conditions in each element) each corresponding to a certain element in the discretization. So in the following we will consider the third Schur complement matrix

$$((-\mathsf{A}/A)/A_{11})/B_{22} = B_{11} - B_{12} B_{22}^{-1} B_{12}^T,$$

induced by the block B_{22} in the matrix $(-\mathsf{A}/A)/A_{11}$. We can prove a similar result to the one given in Theorem 2.1. Therefore, the Schur complement system with the matrix $-(\mathsf{A}/A)/A_{11}$ can be reduced to the Schur complement matrix $((-\mathsf{A}/A)/A_{11})/B_{22}$ of dimension equal to *NIF*, without inducing any additional fill-in. Moreover, this can be done using incomplete factorization procedures.

THEOREM 2.2.

$$Struct(B_{11} - B_{12} B_{22}^{-1} B_{12}^T) = Struct(A_{22}).$$

Proof. Using Theorem 2.1 we get

$$\begin{aligned} Struct(B_{11} - B_{12} B_{22}^{-1} B_{12}^T) &= Struct(B_{11}) + Struct(B_{12} B_{22}^{-1} B_{12}^T) \\ &= Struct(A_{22}) + Struct(A_{23} A_{33}^{-1} A_{23}^T) \\ &= Struct(A_{22}) + Struct(C_1^T A^{-1} C_2 C_2^T A^{-1} C_2 C_2^T A^{-1} C_1). \end{aligned}$$

Since $Struct(C_1^T A^{-1} C_2 C_2^T A^{-1} C_2 C_2^T A^{-1} C_1) \subset Struct(A_{22})$ (with equality only in the trivial singular case with $|\mathcal{E}_h| = 1$ and $NNC = 5$) we get the desired result $Struct(B_{11} - B_{12} B_{22}^{-1} B_{12}^T) = Struct(A_{22})$. □

The following simple corollary gives the number of nonzero elements in the second and third Schur complement matrices $(-\mathsf{A}/A)/A_{11}$ and $((-\mathsf{A}/A)/A_{11})/B_{22}$. We shall use these results later.

COROLLARY 2.1. *The number of nonzeros in the triangular part of* $(-\mathsf{A}/A)/A_{11}$ *is given by*

$$
\begin{aligned}
&|sym((-\mathsf{A}/A)/A_{11})| = \\
&= \frac{1}{2} \sum_{i \in \mathcal{E}_h} (NIF_i + NNC_i)^2 + \frac{1}{2} \sum_{i \in \mathcal{E}_h} \sum_{j \in Adj(i)} NIF_j + \frac{1}{2}(NNC - NIF)
\end{aligned} \tag{2.17}
$$

and the number of nonzeros in the triangular part of $(-\mathsf{A}/A)/A_{11})/B_{22}$ *is given by*

$$
|sym(((-\mathsf{A}/A)/A_{11})/B_{22})| = \frac{1}{2} \sum_{i \in \mathcal{E}_h} NIF_i^2 + \frac{1}{2} \sum_{i \in \mathcal{E}_h} \sum_{j \in Adj(i)} NIF_j - \frac{1}{2} NIF. \tag{2.18}
$$

Apart from explicit assembly of the Schur complement matrices or using them implicitly there is another possibility which may be considered – keeping the Schur complements in factorized form. Consider the following decomposition:

$$
\begin{aligned}
\mathsf{A}/A &= -[L_A^{-1} \begin{pmatrix} B & C_1 & C_2 \end{pmatrix}]^T [L_A^{-1} \begin{pmatrix} B & C_1 & C_2 \end{pmatrix}] \\
&= \begin{pmatrix} \hat{B} & \hat{C}_1 & \hat{C}_2 \end{pmatrix}^T \begin{pmatrix} \hat{B} & \hat{C}_1 & \hat{C}_2 \end{pmatrix},
\end{aligned} \tag{2.19}
$$

where $A = L_A L_A^T$. In contrast to the previous case, where the local numbering of the faces corresponding to the individual elements did not play a role, this is not the case now.

THEOREM 2.3. *Assume that all the elements within the diagonal blocks of the matrix* A *are nonzero. The fill-in in* $\begin{pmatrix} \hat{B} & \hat{C}_1 & \hat{C}_2 \end{pmatrix}$ *is minimal if the faces with Dirichlet boundary conditions are numbered first in the local ordering of each finite element.*

Proof. Because of the block structure of A we can consider the individual finite elements independently. The minimum value of the nonzero count of $\hat{C}_1 \cup \hat{C}_2$ in 5 subsequent rows which correspond to the same finite element is $\frac{1}{2} \sum_{i \in \mathcal{E}_h} (NIF_i + NNC_i)(NIF_i + NNC_i + 1)$ and it is easily checked to be minimal in this case. □

Therefore, from now we assume that within each element we have first numbered the faces corresponding to Dirichlet boundary conditions, then the interior inter-element faces and finally the faces with Neumann boundary conditions. The matrix (2.19) can be written in the form

$$
\begin{pmatrix} \hat{B} & \hat{C}_1 & \hat{C}_2 \end{pmatrix}^T \begin{pmatrix} \hat{B} & \hat{C}_1 & \hat{C}_2 \end{pmatrix} = \begin{pmatrix} \hat{B}^T \hat{B} & \hat{B}^T \hat{C}_1 & \hat{B}^T \hat{C}_2 \\ \hat{C}_1^T \hat{B} & \hat{C}_1^T \hat{C}_1 & \hat{C}_1^T \hat{C}_2 \\ \hat{C}_2^T \hat{B} & \hat{C}_2^T \hat{C}_1 & \hat{C}_2^T \hat{C}_2 \end{pmatrix}. \tag{2.20}
$$

It is clear that it is more advantageous to keep most of the blocks of (2.20) in the explicit form multiplying the factors directly. A typical example is the block $\hat{B}^T \hat{B}$,

which is a diagonal matrix. The main question here is whether we can reduce the system further as in the previous case and at the same time keep the matrix blocks in a factorized form. Unfortunately, there is one basic obstacle. Whereas we are able to embed the structure of $A_{12}^T A_{11}^{-1} A_{12}$ into the structure of A_{22} we cannot in general express $\hat{C}_1^T \hat{C}_1 - C_1^T \hat{B}(\hat{B}^T \hat{B})^{-1} \hat{B}^T C_1 = \hat{C}_1^T (I - \hat{B}(\hat{B}^T \hat{B})^{-1} \hat{B}^T) C_1$ in the factorized form as $\hat{C}_1^T L_{B_{11}} L_{B_{11}}^T \hat{C}_1$, where $L_{B_{11}}$ is factor which can be easily computed.

We have considered the partially factorized structure (2.20) since it is important from a computational point of view. Using a structural prediction based on such factors is exactly the way how to obtain the *sparsity structure* of explicit Schur complement matrices $-\mathsf{A}/A$, $(-\mathsf{A}/A)/A_{11}$ and $((-\mathsf{A}/A)/A_{11})/B_{22}$ in an efficient way. In our implementations we used a procedure similar to the one from [16] to get these structures.

3. Direct solution of the Schur complement systems. In the following we will discuss the direct solution of the Schur complement systems. Namely, we will concentrate on the system with the matrix $((-\mathsf{A}/A)/A_{11})/B_{22} \in \mathcal{R}^{NIF,NIF}$. The following theorem gives a bound on the asymptotic work necessary to solve the linear system (1.12), which is dominated by the decomposition of the matrix $((-\mathsf{A}/A)/A_{11})/B_{22}$.

THEOREM 3.1. *The number of arithmetic operations to solve the symmetric indefinite system (1.12) directly via three consecutive block eliminations and using the Cholesky decomposition is* $O(NIF^2)$.

Proof. We will only give a sketch of the proof here. The work is dominated by the decomposition of $B_{11} - B_{12} B_{22}^{-1} B_{12}^T$, which has the same nonzero structure as A_{22}.

Our uniform regular finite element mesh is a well-shaped mesh in a suitable sense (see [19]). The proof of Lemma 2.2 and the statements of Theorem 2.1 and 2.2 imply that the graph G of $((-\mathsf{A}/A)/A_{11})/B_{22}$ is also the graph of a well-shaped mesh. Namely, it is a bounded-degree subgraph of some overlap graph (see [18], [19]). It was shown in [25] that the upper bound on the second-smallest eigenvalue of the Laplacian matrix of G (the Fiedler value) is of the order $O(1/NIF^{2/3})$. Then using the techniques from [25] we obtain that there exists a $O(NIF^{2/3})$-size bisector of G.

Therefore, G satisfies the so-called $NIF^{2/3}$-separator theorem: there exist constants $1/2 \leq \alpha < 1, \beta > 0$ such that the vertices of G can be partitioned into sets G_A, G_B and the vertex separator G_C such that $|G_A|, |G_B| \leq \alpha NIF$ and $|G_C| \leq \beta NIF^{2/3}$. Moreover, any subgraph of G satisfies the $NIF^{2/3}$-separator theorem. The technique of recursive partitioning of G called generalized nested dissection and used to reorder the considered Schur complement matrix provides an elimination ordering with an $O(NIF^2)$-bound on the arithmetic work of Cholesky decomposition (see Theorem 6 in [12]). ∎

Note that the explicit computation of the matrix $((-\mathsf{A}/A)/A_{11})/B_{22}$ is necessary in the framework of direct methods. Theorem 3.1 provides a theoretical result which is based on spectral partitioning methods. The reordering algorithms based on the separators obtained by the spectral partitioning techniques and applied recursively within the nested dissection need not necessarily be the best practical approach to get a reasonable matrix reordering. Nevertheless, experimental results with various partitioning schemes show

that high quality reorderings can be efficiently computed in this way (see [7]). Also some other reorderings which combine global procedures (partitioning of large meshes) and local algorithms (like MMD) can provide reasonable strategies to find a fill-in minimizing permutation.

4. The conjugate gradient method applied to the Schur complement systems. In this section we concentrate on the iterative solution of the Schur complement systems discussed in Section 3. We consider the conjugate gradient method applied to the symmetric positive definite systems with matrices $-\mathsf{A}/A$, $((-\mathsf{A}/A)/A_{11})$ and $((-\mathsf{A}/A)/A_{11})/B_{22}$. It is well known that the convergence rate of the conjugate gradient method can be bounded in terms of the condition number of the corresponding Schur complement matrix [9], [6], [26]. We show that the condition number of the matrix A/A is asymptotically the same as the conditioning of the negative part of spectrum of the whole indefinite matrix A. Moreover, we prove that condition numbers of the matrices $((-\mathsf{A}/A)/A_{11})$ and $((-\mathsf{A}/A)/A_{11})/B_{22}$ grow like $1/h^2$ with respect to the discretization parameter h and they do not deteriorate during the successive eliminations. Based on these results we estimate the number of iteration steps necessary to achieve the prescribed tolerance in error norm reduction. We show that the number of iteration steps necessary to reduce the error norm by the factor of ε grows asymptotically like $1/h$ for all three Schur complement systems. Therefore, the total number of flops in the iterative algorithm can be significantly reduced due to decrease of the matrix order during the elimination. First, we consider the following theorem.

THEOREM 4.1. *Let $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_{NA} > 0$ be the eigenvalues of the positive definite block $A \in \mathcal{R}^{NA,NA}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{NBC} > 0$ be the singular values of the matrix block $(B\ C) \in \mathcal{R}^{NA,NBC}$. Then for the eigenvalues of the Schur complement matrix $-\mathsf{A}/A = (B\ C)^T A^{-1}(B\ C)$ we have*

$$\sigma(-\mathsf{A}/A) \subset [\sigma^2_{NBC}/\mu_1, \sigma^2_1/\mu_{NA}]. \tag{4.21}$$

Moreover, for the eigenvalues of the positive definite matrix blocks $\begin{pmatrix} A_{22} & A_{23} \\ A_{23}^T & A_{33} \end{pmatrix} = (C_1\ C_2)^T A^{-1}(C_1\ C_2)$ and $A_{11} = B^T A^{-1} B$ it follows

$$\sigma(\begin{pmatrix} A_{22} & A_{23} \\ A_{23}^T & A_{33} \end{pmatrix}) \subset [1/\mu_1, 2/\mu_{NA}], \tag{4.22}$$

$$\sigma(A_{11}) \subset [5/\mu_1, 5/\mu_{NA}]. \tag{4.23}$$

The condition number of the Schur complement system matrix $-\mathsf{A}/A$ then can be bounded by the expression

$$\kappa(-\mathsf{A}/A) \leq \frac{\sigma_1^2 \mu_1}{\sigma^2_{NBC}\mu_{NA}} = \kappa^2((B\ C))\kappa(A). \tag{4.24}$$

Proof. The positive definite matrix A^{-1} has the spectrum $0 < 1/\mu_1 \leq 1/\mu_2 \leq \ldots \leq 1/\mu_{NA}$. The first inclusion in the theorem follows from the following two inequalities

$$\frac{1}{\mu_1}((B\ C)x,(B\ C)x) \leq ((B\ C)^T A^{-1}(B\ C)x,x) \leq \frac{1}{\mu_{NA}}((B\ C)x,(B\ C)x),$$
$$\sigma^2_{NBC}(x,x) \leq ((B\ C)^T(B\ C)x,x) \leq \sigma^2_1(x,x).$$

Similarly, from the inequalities

$$\frac{1}{\mu_1}(Cx,Cx) \leq (C^T A^{-1}Cx,x) \leq \frac{1}{\mu_{NA}}(Cx,Cx),$$
$$(x,x) \leq (C^T Cx,x) \leq 2(x,x)$$

we obtain the second inclusion. The third part of the proof is completely analogous to the second part. □

COROLLARY 4.1. *There exist positive constants c_9 and c_{10} such that for the spectrum of the Schur complement matrix $-\mathsf{A}/A$ we have*

$$\sigma(-\mathsf{A}/A) \subset [c_9/NE, c_{10}/\sqrt[3]{NE}], \tag{4.25}$$

where $c_9 = c_3^2/c_2$ and $c_{10} = c_4^2/c_1$. The condition number of the matrix $-\mathsf{A}/A$ can be bounded as

$$\kappa(-\mathsf{A}/A) \leq c_{11}\sqrt[3]{NE^2}, \quad c_{11} = c_{10}/c_9. \tag{4.26}$$

The Schur complement system with positive definite matrix $-\mathsf{A}/A$ can be solved iteratively by the conjugate gradient method [9] or the conjugate residual method [6]. It is well known that the conjugate gradient method generates the approximate solutions which minimize the energy norm of the error at each iteration step [26], [6]. The closely related conjugate residual method that differ only in the definition of innerproduct, on the other hand, generates the approximate solutions which minimize their residual norm at every iteration [6]. It is also well known fact that there exists so-called peak/plateau connection between these methods [5] showing that there is no significant difference in the convergence rate of these methods when measured by the residual norm of an approximate solution. In our paper we use the conjugate gradient method together with the minimal residual smoothing procedure applied on its top to get monotonic residual norms [28]. Applying such technique allows better monitoring of the convergence by residual norm and it is mathematically equivalent to the residual minimizing conjugate residual method [6]. The computational cost of this technique is minimal and it costs only two inner products and one vector update per iteration. In the framework of iterative methods the number of operations in matrix-vector products is what is usually the most important. These products, performed repeatedly in each iteration loop, contribute in a substantial way to the final efficiency of iterative solver. When solving the system with Schur complement matrix $-\mathsf{A}/A$ the number of flops per iteration for an unpreconditioned method is dominated by the matrix vector multiplication with the matrix $-\mathsf{A}/A$. Its number of nonzeros was given by Lemma 2.2. Moreover, using the

estimates (1.13) and (1.14), the condition number of the Schur complement matrix $-\mathsf{A}/A$ can be bounded by the term $O(\sqrt[3]{NE^2})$. Consequently, the number of flops for conjugate gradients necessary to achieve a reduction by ε is of order

$$O\left(\left[NE + NIF + NNC + \sum_{i\in\mathcal{E}_h}(NIF_i + NNC_i)^2 + \sum_{i\in\mathcal{E}_h}\sum_{j\in Adj(i)} NIF_j\right]\sqrt[3]{NE}\right).$$

Assuming overestimates $(NIF_i + NNC_i) \leq 5$ and $NIF_i \leq 5$, we obtain the asymptotic estimate of order $O(NE\sqrt[3]{NE})$.

The previous considerations did not take into account the Schur complement systems with matrices $((-\mathsf{A}/A)/A_{11})$ and $(((-\mathsf{A}/A)/A_{11})/B_{22})$. The convergence rate of the iterative conjugate gradient method applied to the second and third Schur complement systems depend analogously on the condition number of the Schur complement matrices [9], [6], [26]. The analysis of the spectrum of the matrix $((-\mathsf{A}/A)/A_{11})$ is given in the following theorem.

THEOREM 4.2. *Let $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_{NA} > 0$ be the eigenvalues of the positive definite block $A \in \mathcal{R}^{NA,NA}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{NBC} > 0$ be the singular values of the matrix block $(B\ C) \in \mathcal{R}^{NA,NBC}$. Then for the spectrum of the Schur complement matrix $(-\mathsf{A}/A)/A_{11}$ we have*

$$\sigma((-\mathsf{A}/A)/A_{11}) \subset \left[\sigma^2_{NBC}/\mu_1, 2/\mu_{NA}\right]. \tag{4.27}$$

Consequently, the condition number of the matrix $(-\mathsf{A}/A)/A_{11}$ can be bounded as follows

$$\kappa((-\mathsf{A}/A)/A_{11}) \leq \frac{2}{\sigma_1^2}\kappa(-\mathsf{A}/A). \tag{4.28}$$

Proof. From the definition of the Schur complement matrix $(-\mathsf{A}/A)/A_{11}$ and the statement of Theorem 4.1 we have

$$\begin{aligned}((-\mathsf{A}/A)/A_{11}x, x) &= \left(\begin{pmatrix} A_{22} & A_{23} \\ A_{23}^T & A_{33}\end{pmatrix} x, x\right) - \left(A_{11}^{-1}\begin{pmatrix}A_{12}\ A_{13}\end{pmatrix} x, \begin{pmatrix}A_{12}\ A_{13}\end{pmatrix} x\right) \\ &\leq 2/\mu_{NA}(x,x).\end{aligned}$$

The bound for the minimal eigenvalue can be obtained considering the following result (see [20], p.201):

$$(-\mathsf{A}/A)^{-1} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & A_{22} & A_{23} \\ A_{13}^T & A_{23}^T & A_{33}\end{pmatrix}^{-1} =$$

$$\begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}(A_{12}\ A_{13})[(-\mathsf{A}/A)/A_{11}]^{-1}\begin{pmatrix}A_{12}^T \\ A_{13}^T\end{pmatrix}A_{11}^{-1} & -A_{11}^{-1}(A_{12}\ A_{13})[(-\mathsf{A}/A)/A_{11}]^{-1} \\ -[(-\mathsf{A}/A)/A_{11}]^{-1}\begin{pmatrix}A_{12}^T \\ A_{13}^T\end{pmatrix}A_{11}^{-1} & [(-\mathsf{A}/A)/A_{11}]^{-1}\end{pmatrix}.$$

Then from the interlacing property of the eigenvalue set of symmetric matrix $-\mathsf{A}/A$ (see e.g. [8]) it follows

$$\left\|[(-\mathsf{A}/A)/A_{11}]^{-1}\right\| \leq \left\|(-\mathsf{A}/A)^{-1}\right\| \leq \frac{\mu_1}{\sigma^2_{NBC}}.$$

Considering the previous inequalities we get the lower bound for the minimal eigenvalue of the matrix $(-\mathsf{A}/A)/A_{11}$, which completes the proof. □

We have shown that the condition number of the Schur complement system matrix $(-\mathsf{A}/A)/A_{11}$ is bounded by a multiple of the condition number of the matrix $-\mathsf{A}/A$. Therefore the number of iteration steps for the conjugate gradient method necessary to reduce the error norm(or after smoothing the residual norm) by some factor is asymptotically the same as before. The complexity of the matrix-vector multiplication is lower and according to Corollary 2.1 is of the order

$$O\left(\sum_{i\in\mathcal{E}_h}(NIF_i+NNC_i)^2+\sum_{i\in\mathcal{E}_h}\sum_{j\in Adj(i)}NIF_j+(NNC-NIF)\right).$$

Assuming again the overestimates $(NIF_i + NNC_i) \leq 5$ and $NIF_i \leq 5$, we obtain the asymptotic estimate $O(NE)$. The total number of flops for the conjugate gradients or the conjugate residual method necessary to achieve a reduction by the factor ε is then again of order $O(NE\sqrt[3]{NE})$. From the statements of Theorem 4.1 and Theorem 4.2 it is clear that the reduction to the Schur complement systems does not affect the asymptotic conditioning of the positive definite matrices $-\mathsf{A}/A$ and $(-\mathsf{A}/A)/A_{11}$. The same is true for the spectral properties of the third Schur complement system with the matrix $((-\mathsf{A}/A)/A_{11})/B_{22}$. Since the proof is completely analogous to the proof of Theorem 4.2 we shall present only the following statement (cf. [10], p. 256).

THEOREM 4.3. *The condition number of* $((-\mathsf{A}/A)/A_{11})/B_{22}$ *is bounded by the condition number of the matrix* $(-\mathsf{A}/A)/A_{11}$

$$\kappa(((-\mathsf{A}/A)/A_{11})/B_{22}) \leq \kappa((-\mathsf{A}/A)/A_{11}). \tag{4.29}$$

In the following we present two additional results concerning the the matrix-vector multiplications with Schur complement matrices. Theorem 4.4 compares the number of nonzeros in the Schur complement matrices $(-\mathsf{A}/A)/A_{11}$ and $((-\mathsf{A}/A)/A_{11})/B_{22}$ to the number of nonzeros in the original matrix A.

THEOREM 4.4. *The number of nonzero entries in the matrix* $((-\mathsf{A}/A)/A_{11})$ *or the matrix* $(((-\mathsf{A}/A)/A_{11})/B_{22})$ *is smaller than the number of nonzeros in the matrix* A.

Proof. Using the fact that $2NIF = \sum_{i\in\mathcal{E}_h} NIF_i \leq 5NE$ and also $\sum_{i\in\mathcal{E}_h}(NIF_i + NNC_i) \leq 5NE$, it follows from Lemma 2.1 and Theorem 2.1 that

$$\begin{aligned}
&|(-\mathsf{A}/A)/A_{11}| - |\mathsf{A}| \\
&= \sum_{i\in\mathcal{E}_h} NIF_i^2 - 2NIF + \sum_{i\in\mathcal{E}_h}\sum_{j\in Adj(i)} NIF_j + \sum_{i\in\mathcal{E}_h} NNC_i^2 \\
&+2\sum_{i\in\mathcal{E}_h} NIF_i NNC_i - 35NE - 4NIF - 2NNC \\
&= \sum_{i\in\mathcal{E}_h}(NIF_i+NNC_i)^2 - 5\sum_{i\in\mathcal{E}_h} NIF_i - 4\sum_{i\in\mathcal{E}_h} NNC_i - \sum_{i\in\mathcal{E}_h}\sum_{j\in Adj(i)} NNC_j \\
&+2\sum_{i\in\mathcal{E}_h}(NIF_i+NNC_i) + \sum_{i\in\mathcal{E}_h}\sum_{j\in Adj(i)}(NIF_j+NNC_j) - 35NE
\end{aligned}$$

$$\leq 2 \sum_{i \in \mathcal{E}_h} (NIF_i + NNC_i) + \sum_{i \in \mathcal{E}_h} \sum_{j \in Adj(i)} (NIF_j + NNC_j) - 35NE \leq 0.$$

Clearly, the number of nonzeros in the matrix $(((-\mathsf{A}/A)/A_{11})/B_{22})$ is even smaller. □

Note that the number of nonzeros in the original matrix A can be smaller or larger than the corresponding number of nonzeros in the matrix $-\mathsf{A}/A$. Consider now the factorized Schur complement in the form (2.20). It can be shown also that there is no clear winner between the number of floating-point operations to multiply a dense vector by the matrix $(\,\hat{B} \quad \hat{C}_1 \quad \hat{C}_2\,)^T\,(\,\hat{B} \quad \hat{C}_1 \quad \hat{C}_2\,)$ or the number of operations to get a product of a matrix $((-\mathsf{A}/A)/A_{11})/B_{22}$ with a dense vector of appropriate dimension, respectively. The result depends on the shape of the domain and its boundary conditions. Nevertheless, the following Theorem 4.5 shows that if we do not form the Schur complement explicitly it is worth to use the factorized form (2.19) and the reordering of the Schur complement from Theorem 2.3 instead of its implicit form.

THEOREM 4.5. *Let v be a dense vector. The number of floating-point operations to compute $(\,\hat{B} \quad \hat{C}_1 \quad \hat{C}_2\,)^T\,(\,\hat{B} \quad \hat{C}_1 \quad \hat{C}_2\,)\,v$ is smaller than the number of floating-point operations to compute*

$$\begin{pmatrix} B^T \\ C_1^T \\ C_2^T \end{pmatrix} A^{-1}\,(\,B \quad C_1 \quad C_2\,)\,v.$$

Proof. Taking into account the local ordering from Theorem 2.3 the difference between these two quantities can be bounded by

$$\begin{aligned} &10NE + \sum_{i \in \mathcal{E}_h} (NIF_i + NNC_i)(NIF_i + NNC_i + 1) - 35NE - 4NIF - 2NNC \\ &\leq -2NIF - NNC \leq 0. \end{aligned}$$

□

5. Numerical experiments. In the following we present numerical experiments which illustrate the results developed in the theoretical part of the paper.

Two model potential flow problems (1.1) and (1.2) in a rectangular domain with Neumann conditions prescribed on the bottom and on the top of the domain have been considered. Dirichlet conditions that preserve the nonsingularity of the whole system matrix A were imposed on the rest of the boundary. The choice of boundary conditions in these examples is motivated by our application and it comes from a modelling of a confined aquifer (see [3]) between two impermeable layers.

In order to verify the theoretical results derived in previous sections we will restrict our attention first to the simplest geometrical shape - cubic domain and report the results obtained from a uniformly regular mesh refinement. In practical situations, however, relatively thin aquifers with possible cracks in the rock are frequently modelled, and so the number of Neumann conditions may represent a big portion of the whole boundary. As our second model example, we consider a rectangular domain discretized by 6 layers of

TABLE 1

Model potential fluid flow problem - cubic domain

Discretization parameters			Matrix dimensions			
h, NE	NIF	NNC	A	$-\mathsf{A}/A$	$(-\mathsf{A}/A)/A_{11}$	$((-\mathsf{A}/A)/A_{11})/B_{22}$
1/5, 250	525	100	2125	875	625	525
1/10, 2000	4600	400	17000	7000	5000	4600
1/15, 6750	15975	900	57375	23625	16875	15975
1/20, 16000	38400	1600	136000	56000	40000	38400
1/25, 31250	75625	2500	265625	109375	78125	75625
1/30, 54000	131400	3600	459000	189000	135000	131400
1/35, 87750	209475	4900	728875	300125	214375	209475
1/40, 128000	313600	6400	1088000	448000	320000	313600

TABLE 2

Model potential fluid flow problem - realistic domain

Discretization parameter			Matrix dimension			
NE	NIF	NNC	A	$-\mathsf{A}/A$	$(-\mathsf{A}/A)/A_{11}$	$((-\mathsf{A}/A)/A_{11})/B_{22}$
35x35x6	33880	4900	126980	53480	38780	33880
45x45x6	56160	8100	210060	88560	64260	56160
55x55x6	84040	12100	313940	132440	96140	84040
65x65x6	117520	16900	438620	185120	134420	117520
75x75x6	156600	22500	584100	246600	179100	156600
85x85x6	201280	28900	750380	316880	230180	201280
95x95x6	251560	36100	937460	395960	287660	251560
105x105x6	307440	44100	1145340	483840	351540	307440

elements in the mesh. As we will see later, the reduction to the third Schur complement proposed in this paper can become even more significant than for the cubic domain. Prismatic discretizations of domains with NE elements were used [14], [11]. For the cubic domain we have then $NE = 2/h^3$. Discretization parameters h, NE, NIF, NNC, dimension N of the resulting indefinite system matrix A and the dimensions of the corresponding Schur complement matrices $-\mathsf{A}/A$, $(-\mathsf{A}/A)/A_{11}$ and $((-\mathsf{A}/A)/A_{11})/B_{22}$ are given in Table 1 for a cubic domain and in Table 2 for a more realistic domain. We note again that the difference between dimensions of the second and third Schur complement matrix is significantly larger in the case of modelling of thin layers that arise regularly in our application.

For the example of a cubic domain the spectral properties of the matrix blocks A and $(B\ C)$ as well as of the whole symmetric indefinite matrix A have been investigated. The extremal positive and negative eigenvalues of the matrix A and the extremal singular values of the block $(B\ C)$ (squared roots of the extremal eigenvalues of the matrix $(B\ C)^T(B\ C)$) were approximated by a reduction to the symmetric tridiagonal form of the matrix using 1500 steps of the symmetric Lanczos algorithm [8] and by a subsequent

TABLE 3

Spectral properties of the system matrix and its blocks - problem with a cubic domain

	matrix blocks spectral properties		eigenvalues of the matrix A	
NE	spectrum of A	sing. values of $(B\|C)$	negative part	positive part
250	[0.16e-2, 0.1e-1]	[0.181e0, 2.63]	[-2.63 , -0.180e0]	[0.166e-2, 2.63]
2000	[0.33e-2, 0.2e-1]	[0.927e-1, 2.64]	[-2.64, -0.898e-1]	[0.335e-2, 2.64]
6750	[0.50e-2, 0.3e-1]	[0.622e-1, 2.64]	[-2.64, -0.354e-1]	[0.509e-2, 2.65]
16000	[0.66e-2, 0.4e-1]	[0.467e-1, 2.64]	[-2.64, -0.413e-1]	[0.679e-2, 2.65]
31250	[0.83e-2, 0.5e-1]	[0.374e-1, 2.65]	[-2.64, -0.311e-1]	[0.861e-2, 2.65]
54000	[0.99e-2, 0.6e-1]	[0.312e-1, 2.65]	[-2.64, -0.241e-1]	[0.104e-1, 2.65]
87750	[0.11e-1, 0.7e-1]	[0.268e-1, 2.65]	[-2.64, -0.190e-1]	[0.120e-1, 2.65]
128000	[0.13e-1, 0.8e-1]	[0.234e-1, 2.65]	[-2.64, -0.152e-1]	[0.136e-1, 2.65]

eigenvalue computation of the resulting tridiagonal matrix using the LAPACK double precision subroutine DSYEV [1]. Extremal eigenvalues of the diagonal matrix block A were computed directly by the LAPACK eigenvalue solver element by element. It can be seen that the computed extremal eigenvalues of the block A are in perfect agreement with the theory (see Table 3). Similarly, we can observe approximately a linear decrease of the computed minimal singular value of the matrix block $(B\ C)$ with respect to the mesh discretization parameter h. From the computed extremal eigenvalues of the whole indefinite system A we can conclude that even if our mesh size parameters h are rather small and give rise to very large system dimensions (see Table 1), they are outside of the asymptotic inclusion set (1.15). Indeed for our example and our mesh size interval we have $c_1/h \ll c_4$, $c_2/h \ll c_4$ and with the exception of $h = 1/35$ and $h = 1/40$ also $c_2/h < c_3 h$. Then using Lemma 2.1 in [22], pp. 3-4 (see also [15]) we obtain the inclusion set in the form

$$\sigma(\mathsf{A}) \subset [-c_4, -\frac{1}{2}(c_2\sqrt[3]{NE} - \sqrt{(c_2\sqrt[3]{NE})^2 + 4(c_3/\sqrt[3]{NE})^2})] \cup [c_1\sqrt[3]{NE}, c_4], \tag{5.30}$$

which is in good agreement with the results in Table 3.

Using the same technique we have approximated the extremal eigenvalues of the Schur complement matrices $-\mathsf{A}/A$, $(-\mathsf{A}/A)/A_{11}$ and $((-\mathsf{A}/A)/A_{11})/B_{22}$ coming from a problem on a cubic domain. From Table 4 it can be seen that the inclusion set for the extremal eigenvalues of the first Schur complement matrix $-\mathsf{A}/A$ coincides with the bounds given in Theorem 4.1. We can see that the extremal eigenvalues of the second Schur complement matrix $(-\mathsf{A}/A)/A_{11}$ are bounded by the extremal eigenvalues of the matrix $-\mathsf{A}/A$. Similarly, the extremal eigenvalues of the third Schur complement matrix $((-\mathsf{A}/A)/A_{11})/B_{22}$ are bounded by the extremal eigenvalues of the matrix $(-\mathsf{A}/A)/A_{11}$. This behaviour is in accordance with the asymptotic bounds given in Theorem 4.2 and Theorem 4.3.

The smoothed conjugate gradient method has been applied to the resulting three Schur complement systems (see also the discussion in previous section). Unpreconditioned and also preconditioned versions with the IC(0) preconditioner [23], [24] for the solution of these symmetric positive definite systems have been used. For the solution of

TABLE 4

Spectral properties of Schur complement matrices - problem with a cubic domain

	spectral properties of Schur complement matrices		
NE	$-\mathsf{A}/A$	$(-\mathsf{A}/A)/A_{11}$	$((-\mathsf{A}/A)/A_{11})/B_{22}$
250	[0.138e2, 0.343e4]	[0.187e2, 0.117e4]	[0.220e2, 0.117e4]
2000	[0.182e1, 0.173e4]	[0.251e1, 0.596e3]	[0.272e1, 0.596e3]
6750	[0.547e0, 0.115e4]	[0.760e0, 0.399e3]	[0.801e0, 0.399e3]
16000	[0.232e0, 0.868e3]	[0.323e0, 0.299e3]	[0.336e0, 0.299e3]
31250	[0.119e0, 0.694e3]	[0.166e0, 0.239e3]	[0.171e0, 0.239e3]
54000	[0.693e-1, 0.579e3]	[0.966e-1, 0.199e3]	[0.992e-1, 0.199e3]
87750	[0.437e-1, 0.496e3]	[0.610e-1, 0.171e3]	[0.624e-1, 0.171e3]
128000	[0.293e-1, 0.434e3]	[0.409e-1, 0.149e3]	[0.417e-1, 0.149e3]

the whole indefinite system the minimal residual method has been used. For the preconditioned version the positive definite block-diagonal preconditioning with ILUT(0,20) for the decomposition of the block corresponding to constraints (see e.g. [22], [21]) was used. The choice of ILUT(0,20) was motivated by our effort to obtain rather precise factorization with restricted memory requirements which should be close to the full decomposition of the block $(B\ C)^T(B\ C)$. This preconditioner was found generally better than the indefinite block-diagonal preconditioning with the same ILUT(0,20) decomposition or than the indefinite preconditioner discussed in [13] or [21]. The initial approximation x_0 was set to zero, the relative residual norm $\frac{\|r_n\|}{\|r_0\|} = 10^{-8}$ was used as the stopping criterion. For the implementation details of iterative solvers we refer to [6]. Our experiments were performed on an SGI Origin 200 with processor R10000. In Table 5 and Table 6 we consider iteration counts and CPU times in the minimal residual method (unpreconditioned/preconditioned) applied to the whole system (1.12) and in the conjugate gradient method (unpreconditioned/preconditioned) applied to the Schur complement systems with the matrices $-\mathsf{A}/A$, $(-\mathsf{A}/A)/A_{11}$ and $((-\mathsf{A}/A)/A_{11})/B_{22}$ for a model problem with a cubic and more realistic domain, respectively. The dependence of the iteration counts presented in all columns of Table 5 corresponds surprisingly well to the theoretical order $O(\sqrt[3]{NE})$. The convergence behaviour of the smoothed conjugate gradient method applied to the third Schur complement system with the matrix $((-\mathsf{A}/A)/A_{11})/B_{22}$ for this case is presented in Figure 4. From the results in Table 5 and Table 6 it follows that while the gain from the solution of the third Schur complement system is rather moderate in the case of a cubic domain and in the case of the realistic flat domain it becomes more significant.

6. Conclusions. Successive block Schur complement reduction for the solution of symmetric indefinite systems has been considered in the paper. It was shown that due to the particular structure of matrices which arise from mixed-hybrid finite element discretization of the potential fluid flow problem, the resulting Schur complement matrices remain sparse. Moreover, their spectral properties do not deteriorate and the iterative conjugate gradient method can be successfully applied. Theoretical bounds for the

TABLE 5

Number of iterations of the conjugate gradient method - problem with a cubic domain

NE	unpreconditioned/preconditioned CG applied to matrix			
	A	$-\mathsf{A}/A$	$(-\mathsf{A}/A)/A_{11}$	$((-\mathsf{A}/A)/A_{11})/B_{22}$
250	319/44	82/20	48/18	43/18
	0.41/0.09	0.05/0.02	0.02/0.01	0.02/0.01
2000	608/76	154/35	87/32	80/32
	6.43/1.56	0.76/0.25	0.30/0.16	0.25/0.14
6750	867/113	223/51	126/48	118/48
	48.17/11.91	3.86/1.50	1.51/0.92	1.30/1.01
16000	1031/138	288/67	164/63	155/63
	161.75/38.86	16.07/5.93	5.59/3.87	5.02/3.43
31250	1195/165	353/82	199/78	192/78
	410.45/100.49	45.69/16.75	16.31/9.94	15.13/9.60
54000	1358/188	418/95	234/93	228/93
	926.98/218.78	104.76/36.92	39.88/24.04	37.94/23.13
85750	1503/205	482/108	269/108	263/108
	1694.68/396.03	216.10/74.42	76.21/47.55	72.00/45.71
128000	1637/229	546/122	303/122	298/122
	2825.94/675.49	389.37/133.10	143.28/87.63	138.72/87.28

TABLE 6

Number of iterations of the conjugate gradient method - realistic model example

NE	unpreconditioned/preconditioned CG applied to matrix			
	A	$-\mathsf{A}/A$	$(-\mathsf{A}/A)/A_{11}$	$((-\mathsf{A}/A)/A_{11})/B_{22}$
14700	1688/209	444/97	248/93	233/93
	231.51/54.57	22.96/8.12	7.52/5.27	6.89/4.35
24300	2043/265	571/122	316/117	296/117
	510.65/123.32	54.46/18.48	20.51/11.69	16.44/10.30
36300	2225/300	692/147	382/142	359/142
	919.43/224.74	118.83/37.24	38.93/22.75	31.16/19.57
50700	2053/336	810/172	448/166	421/166
	1547.14/370.18	195.08/62.90	73.65/40.68	57.01/34.15
67500	2723/365	927/197	513/190	482/190
	2343.57/559.24	316.33/103.93	116.30/65.70	93.15/56.55
86700	2959/403	1042/222	578/214	543/214
	3374.38/814.61	495.97/160.23	175.55/99.90	138.81/83.85
108300	3211/429	1256/247	741/254	741/255
	4621.81/1086.38	821.96/240.74	306.28/158.08	253.10/132.33
132300	3420/447	1272/271	706/262	663/262
	6012.54/1382.56	1009.94/323.72	374.90/208.07	299.87/177.34

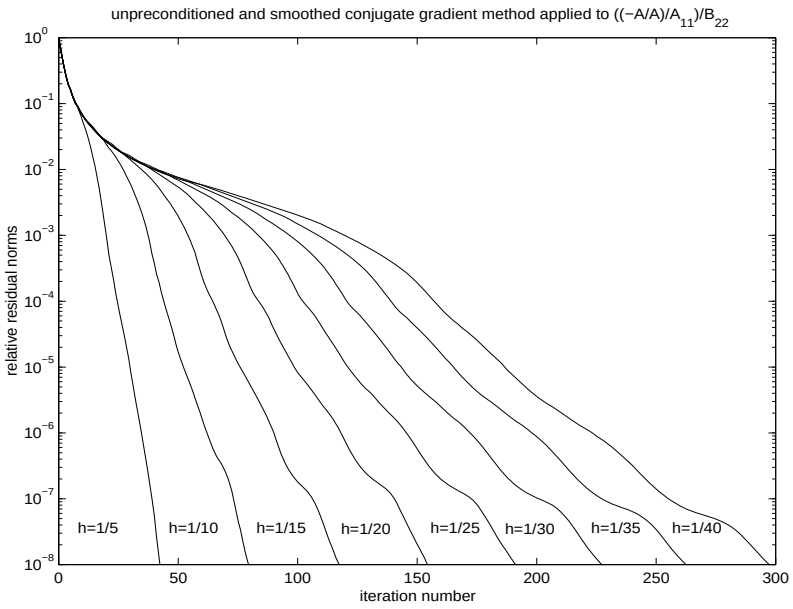

FIG. 4. *Convergence of the smoothed conjugate gradient method applied to the third Schur complement system*

convergence rate of this method in terms of the discretization parameters have been developed and tested on a model problem example. Numerical experiments indicate that the given theoretical bounds on the eigenvalue set are realistic not only for the system matrix and its blocks, but also for the Schur complement matrices. The iteration counts for the conjugate gradient method are also in a good agreement with the theoretical predictions. Direct solution of the third Schur complement system is also a possible alternative. Nevertheless, its comparison with iterative solvers is outside the scope of this paper.

In case of structured grids, a geometric multigrid solver and/or preconditioner for solving the final Schur complement system can be used. Namely, the stencil from the first Schur complement which expresses element-element connectivity in the domain (see proof of Lemma 2.2) remains unchanged after the subsequent two reduction and an appropriate method could be based on that.

Another approach for the solution of symmetric indefinite systems seems to be promising. As was pointed out in [2], the classical null-space algorithm can be implemented. QR factorization of the off-diagonal block $(B\ C)$ is considered and the solution of the indefinite system is transformed to the solution of a block lower triangular system, where the subproblem corresponding to the diagonal block can be solved using the Cholesky factorization or an iterative conjugate gradient-type algorithm. This approach has the advantage of performing the matrix-vector multiplication by the Q factor using elementary Householder transformations. Although the Q factor may be structurally full, the elementary Householder vectors may be quite sparse. Moreover, a roundoff error analysis of the algorithm can be carried out.

7. Acknowledgment. Authors would like to thank Michele Benzi for careful reading of manuscript and anonymous referees for their many useful comments which significantly improved the presentation of the paper. We are indebted to Jiří Mužák from the Department of Mathematical Modelling in DIAMO, s.e., Stráž pod Ralskem for providing us with a model numerical example for the experimental part of this paper and to Jörg Liesen for giving us the reference [20]. This work was supported by the Grant Agency of the Czech Republic under grant 201/98/P108 and by the grant AS CR A2030706.

REFERENCES

[1] E. Anderson et al. LAPACK User's Guide *SIAM, Philadelphia, 1992.*

[2] M. Arioli. The use of QR factorization in sparse quadratic programming. *Tech. Rep. 1070, Istituto di Analisi Numerica del C.N.R., Pavia, Italy, 1998.*

[3] J. Bear. Dynamics of Fluids in Porous Media. *American Elsevier Pub. Company, New York, 1972.*

[4] F. Brezzi and M. Fortin. Mixed and Hybrid Finite Element Methods. *Springer-Verlag, 1991.*

[5] J. Cullum and A. Greenbaum. Relations between Galerkin and norm-minimizing iterative methods for solving linear systems. *SIAM J. Matrix Anal. Appl., 17 (1996), pp. 223-247.*

[6] H.C. Elman. Iterative methods for large, sparse, nonsymmetric systems of linear equations. *Research Report No. 229, Yale University, 1982.*

[7] J.R. Gilbert, G.L. Miller and S.H. Teng. Geometric mesh partitioning: implementation and experiments. *Proc. 9th International Parallel Proc. Symposium, IEEE Computer Society Press, 418 – 427, 1995.*

[8] G.H. Golub and C.F. van Loan. Matrix Computations. 2nd edition. *The Johns Hopkins University Press, Baltimore, 1989.*

[9] M.R. Hestenes and E. Stiefel. Method of conjugate gradients for solving linear systems. *J. Res. Nat. Bureau Standards, 49(1952), 409–435.*

[10] N.J. Higham. Accuracy and Stability of Numerical Algorithms. *SIAM, Philadelphia, 1996.*

[11] E.F. Kaasschieter and A.J.M. Huijben. Mixed-hybrid finite elements and streamline computation for the potential flow problem, *Numerical Methods for Partial Differential Equations 8 (1992), 221-266.*

[12] R.J. Lipton, D.J. Rose and R.J. Tarjan. Generalized nested dissection, *SIAM J. Numer. Anal., 16 (1979), 346–358.*

[13] L. Lukšan and J. Vlček. Indefinitely preconditioned inexact Newton method for large sparse equality constrained non-linear programming problems. *Num. Linear Algebra with Appl. 5 (1998), 219–247.*

[14] J. Maryška, M. Rozložník and M. Tůma. Mixed-hybrid finite element approximation of the potential fluid flow problem. *J. Comput. Appl. Math. 63 (1995), 383-392.*

[15] J. Maryška, M. Rozložník and M. Tůma. The potential fluid flow problem and the convergence rate of the minimal residual method. *Numer. Linear Algebra with Applications 3(6) (1996), 525-542.*

[16] P. Matstoms. Sparse QR factorization with applications to linear least squares problems. *Linköping Studies in Science and Technology. Dissertations. No. 337, Department of Mathematics, Linköping University, Linköping, 1994.*

[17] J.D. Moulton, J.E. Morel and U.M. Ascher. Approximate Schur complement preconditioning of the lowest-order nodal discretizations. *SIAM J. Sci. Comput. 19 (1998), 185–205.*

[18] G.L. Miller, S.H. Teng, W. Thurston and S.A. Vavasis. Automatic mesh partitioning, *in Graph Theory and Sparse Matrix Computation, A. George, J. Gilbert nd J. Liu, eds., vol. 56 of IMA Volumes in Mathematics and Its Applications, Springer-Verlag, 1993, 57–84.*

[19] G.L. Miller, S.H. Teng, W. Thurston and S.A. Vavasis. Geometric separators for finite element meshes. *SIAM J. Sci. Comput. 19 (1998), 364–396.*

[20] D.V. Ouellette. Schur Complement and Statistics. *Linear Algebra and Its Applications 36, 1981, 187–295.*

[21] I. Perugia and V. Simoncini. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. See also An optimal indefinite preconditioner for a mixed finite element method. *Tech. Report 1098, IAN CNR, Pavia, Italy, 1998.*

[22] T. Rusten and R. Winther. A preconditioned iterative method for saddle point problems. *SIAM J. Math. Anal. Appl. 13, 1992, 887–905.*

[23] Y. Saad. Iterative Methods for Sparse Linear Systems. *PWS Publishing Company, ITP, 1996.*

[24] Y. Saad. ILUT: A dual threshold incomplete ILU factorization. *Num. Lin. Alg. Appl.*, 1:387–402, 1994.

[25] D.A. Spielman and S.H. Teng. Spectral partitioning works: Planar graphs and finite element meshes. *Technical Report UCB//CSD-96-898, University of Berkeley, 1996.*

[26] H. van der Vorst. Parallel iterative solution methods for linear systems arising from discretized PDE's. *preprint, Univerity of Utrecht, 1996.*

[27] C. Wagner, W. Kinzelbach and G. Wittum. Schur-complement multigrid: A robust method for groundwater flow and transport problems. *Num. Math. 75 (1997), 523–545.*

[28] R. Weiss. Parameter-Free Iterative Linear Solvers *Mathematical Research Series Vol. 97, Akademie Verlag, Berlin, 1996.*