Block preconditioning strategies for high order finite element discretization of the time-harmonic Maxwell equations

Matthias Bollhoefer\textsuperscript{1} and Stéphane Lanteri\textsuperscript{2}

Abstract We study block preconditioning strategies for the solution of large sparse complex coefficients linear systems resulting from the discretization of the time-harmonic Maxwell equations by a high order discontinuous finite element method formulated on unstructured simplicial meshes. The proposed strategies are based on principles from incomplete factorization methods. Moreover, a complex shift is applied to the diagonal entries of the underlying matrices, a technique that has recently been exploited successfully in similar contexts and in particular for the multigrid solution of the scalar Helmholtz equation. Numerical results are presented for 2D and 3D electromagnetic wave propagation problems in homogeneous and heterogeneous media.

1 Introduction

The present study is concerned with the development of a high-performance numerical methodology for the computer simulation of time-harmonic electromagnetic wave propagation problems in irregularly shaped domains and heterogeneous media. In this context, we are naturally led to consider volume discretization methods (\textit{i.e.} finite difference, finite volume or finite element methods) as opposed to surface discretization methods (\textit{i.e.} boundary element method). Most of the related existing works deal with the second-order form of the time-harmonic Maxwell equations discretized by a conforming finite element method [13]. More recently, discontinuous Galerkin (DG) methods have also been considered for this purpose. Here, we con-
centrate on the first-order form of the time-harmonic Maxwell equations discretized by a high order discontinuous Galerkin method formulated on unstructured simplicial meshes. While it keeps almost all the advantages of the finite element method (large spectrum of applications, complex geometries, etc.), the DG method has other nice properties which explain the renewed interest it gains in various domains in scientific computing (as witnessed by books or special issues of journals dedicated to this method [4–6]): easy extension to higher order interpolation (one may increase the degree of the polynomials in the whole mesh as easily as for spectral methods and moreover this can be done very locally), no global mass matrix to invert (when solving time-domain systems of partial differential equations using an explicit time scheme), easy handling of complex meshes (the grid may be a classical conforming finite element mesh, a non-conforming one or even a hybrid mesh made of various types of element), natural treatment of discontinuous solutions and coefficient heterogeneities, nice parallelization properties (the compact nature of a discontinuous Galerkin scheme is in favor of high computation to communication ratio especially for high order interpolation methods).

The DG discretization of the first order form of the time-harmonic Maxwell equations leads to a large sparse complex system of equations that exhibits a block structure which is linked to the use of a polynomial interpolation method for the approximation of the electromagnetic field within a mesh element. For moderately large 2D problems, this system can be efficiently solved by an optimized sparse solver such as MUMPS [1]. However, for large 2D problems or for 3D problems, such a solution strategy is simply not feasible. In [8], a hybrid iterative-direct solver is proposed for the solution of the linear system resulting from the DG discretization of the 3D time-harmonic Maxwell equations. The starting-point is a Schwarz type domain decomposition algorithm which is defined for the continuous time-harmonic Maxwell equations and which is based on appropriate transmission conditions between neighboring subdomains. At the discrete level, this domain decomposition solver combines an iterative solver acting on a reduced linear system of equations involving interface unknowns, with a sparse direct solver within each subdomain. For moderately large 3D problems and for the lowest interpolation degrees (i.e. 0-th and 1-st order) in the DG method, the resulting hybrid iterative-direct solver is a viable solution strategy. However, for very large problems and for high interpolation degrees, the size of the subdomain problems prohibits the use of a sparse direct solver. Besides, increasing the number of subdomains to reduce the size of the local problems is generally not a proper approach since this incurs numerical scalability issues which have not been investigated so far for optimized Schwarz methods.

In this paper we will discuss an alternative way of solving the discretized time-harmonic Maxwell equations. Our approach is mainly based on the relations between the second-order Maxwell equations and Helmholtz equations. For Helmholtz equations, recently numerical methods have been presented that are based on the shifted Laplacian [2, 9, 12, 14]. I.e., first an artificial damping is introduced into the equations which results in an additional imaginary shift. Then the numerical approximation is computed for the shifted system instead of the original system. Finally, the approximation is applied to the original equations. For the first order
time-harmonic Maxwell equations an analogous perturbation is performed that implicitly shifts the second order systems. The numerical approximation we apply to the shifted system is based on a multilevel block incomplete factorization that uses a pivoting strategy to deal with small pivots. Furthermore, our block factorization approach is designed to deal with large blocks in order to preserve the natural block structure which is obtained from the discontinuous Galerkin discretization. Numerical experiments confirm that this approach is able to efficiently solve the time-harmonic Maxwell equations.

2 The continuous problem

We consider solving the normalized time-harmonic Maxwell equations in the first order form:

\[ i \omega \varepsilon_r E - \text{curl} H = -J, \quad i \omega \mu_r H + \text{curl} E = 0, \]

where \( E \) and \( H \) are the unknown electric and magnetic fields and \( J \) is a known current source; \( \varepsilon_r \) and \( \mu_r \) respectively denote the relative electric permittivity and the relative magnetic permeability and we assume here the case of a linear isotropic non-magnetic (i.e. \( \mu_r = 1 \)) media. The relative electric permittivity is linked to its absolute value through \( \varepsilon = \varepsilon_r \varepsilon_0 \) where \( \varepsilon_0 \) is the permittivity of the vacuum. The angular frequency of the problem is given by \( \omega \). Equations (1) are solved in a bounded domain \( \Omega \). On the boundary \( \partial \Omega = \Gamma_m \cup \Gamma_a \), the following boundary conditions are imposed:

- a perfect electric conductor (PEC) condition on \( \Gamma_m : n \times E = 0 \),
- a Silver-Müller absorbing condition on \( \Gamma_a : \mathcal{L}(E, H) = \mathcal{L}(E^{\text{inc}}, H^{\text{inc}}) \),

where \( \mathcal{L}(E, H) = n \times E - Zn \times (H \times n) \) with \( Z = \sqrt{\mu_r/\varepsilon_r} \). The vectors \( E^{\text{inc}} \) and \( H^{\text{inc}} \) represent the components of an incident electromagnetic wave and \( n \) denotes the unit outward normal. Equations (1) and (2) can be further rewritten in the form:

\[
\begin{cases}
    i \omega G_0 W + G_l \partial_l W + G_x \partial_x W + G_z \partial_z W = -J \text{ in } \Omega, \\
    (M_{\Gamma_m} - G_n) W = 0 \text{ on } \Gamma_m, \\
    (M_{\Gamma_a} - G_n)(W - W^{\text{inc}}) = 0 \text{ on } \Gamma_a,
\end{cases}
\]

where \( W = (E, H)^T \) is the new unknown vector, \( J = (J_x, 0)^T \) and:

\[
G_0 = \begin{pmatrix} \varepsilon_r I_3 & 0_3 \\ 0_3 & \mu_r I_3 \end{pmatrix}, \quad G_l = \begin{pmatrix} 0_3 & N_{l'} \\ N^T_{l'} & 0_3 \end{pmatrix}, \quad N_v = \begin{pmatrix} 0 & v_x & -v_y \\ -v_z & 0 & v_x \\ v_y & -v_x & 0 \end{pmatrix},
\]

with the index set \( l \in \{ x, y, z \} \) for \( G_l \) and where \( (e_x, e_y, e_z) \) is the canonical basis of \( \mathbb{R}^3 \) and \( v = (v_x, v_y, v_z)^T \). The term \( I_3 \) denotes the identity matrix, and \( 0_3 \) the null ma-
tion of the discrete problem which aims at finding most special elements \( K \). Let
\[
\Omega
\]
its imaginary part, which appears for instance in the case of conductive materials, is only gathers the positive (respectively negative) eigenvalues.
\[
\Gamma
\]
oriented face \( F \) and elements whose entries are equal to 0 if the face \( F \) and their orientations match, and -1 if \( F \) and their orientations do not match. For \( F = \partial K \cap \partial \bar{K} \), we also define \( \{ \mathbf{V} \} = I_{FK} \mathbf{V}_{|K} + I_{FK} \mathbf{V}_{|\bar{K}} \) and \( \{ \nabla \} = \frac{1}{2} \left( \mathbf{V}_{|K} + \mathbf{V}_{|\bar{K}} \right) \). Finally, the matrix \( S_F \), which is hermitian positive semi-definite, permits to penalize the jump of a field or some components of this field on the face \( F \), and the matrix \( M_{FK} \) insures the asymptotic consistency with the boundary conditions of the continuous problem. Problem (4) is often interpreted in terms of local problems in each element \( K \) of \( \Omega_h \) coupled by the introduction of an element
\[
M_{\Gamma_0} = \begin{pmatrix} 0_3 & N_0 \\ -N_0 & 0_3 \end{pmatrix} \quad \text{and} \quad M_{\Gamma_m} = |G_n|.
\]

3 Discretization by a discontinuous Galerkin method

Let \( \Omega_h \) denote a discretization of the domain \( \Omega \) into a union of conforming simplicial elements \( K \). We look for the approximate solution \( \mathbf{W}_h \) of (3) in \( V_h \times V_h \) where the functional space \( V_h \) is defined by \( V_h = \{ \mathbf{U} \in L^2(\Omega) \}^3 / \forall K \in \Omega_h, \mathbf{U}|_K \in P_p(K) \} \), where \( P_p(K) \) denotes a space of vectors with polynomial components of degree at most \( p \) over the element \( K \). The DG discretization of system (3) yields the formulation of the discrete problem which aims at finding \( \mathbf{W}_h \) in \( V_h \times V_h \) such that:

\[
\begin{align}
\int_{\Omega_h} (i \omega G_0 \mathbf{W}_h)^T \mathbf{\nabla} \mathbf{d}v &+ \sum_{K \in \Omega_h} \int_K \left( \sum_{l \in \{x,y,z\}} G_l \partial_l(\mathbf{W}_h) \right)^T \mathbf{\nabla} \mathbf{d}v \\
+ \sum_{F \in \Gamma_0} \int_F \left( \frac{1}{2} (M_{FK} - I_{FK} G_{nf}) \mathbf{W}_h \right)^T \mathbf{\nabla} \mathbf{d}s &- \sum_{F \in \Gamma_0} \int_F (G_{nf} \{ \mathbf{W}_h \})^T \{ \mathbf{\nabla} \} \mathbf{d}s + \sum_{F \in \Gamma_0} \int_F (S_F \{ \mathbf{W}_h \})^T \{ \mathbf{\nabla} \} \mathbf{d}s
\end{align}

\]

where \( \Gamma_0, \Gamma^u \) and \( \Gamma^m \) respectively denote the set of interior (triangular) faces, the set of faces on \( \Gamma_0 \) and the set of faces on \( \Gamma_m \). The unitary normal associated with the oriented face \( F \) is \( n_F \) and \( I_{FK} \) stands for the incidence matrix between oriented faces and elements whose entries are equal to 0 if the face \( F \) does not belong to element \( K, 1 \) if \( F \in K \) and their orientations match, and -1 if \( F \in K \) and their orientations do not match. For \( F = \partial K \cap \partial \bar{K} \), we also define \( \{ \mathbf{V} \} = I_{FK} \mathbf{V}_{|K} + I_{FK} \mathbf{V}_{|\bar{K}} \) and \( \{ \mathbf{\nabla} \} = \frac{1}{2} \left( \mathbf{V}_{|K} + \mathbf{V}_{|\bar{K}} \right) \). Finally, the matrix \( S_F \), which is hermitian positive semi-definite, permits to penalize the jump of a field or some components of this field on the face \( F \), and the matrix \( M_{FK} \) insures the asymptotic consistency with the boundary conditions of the continuous problem. Problem (4) is often interpreted in terms of local problems in each element \( K \) of \( \Omega_h \) coupled by the introduction of an element

\footnote{If \( T A T^{-1} \) is the eigendecomposition of \( G_n \), then \( G_n^+ = T A^+ T^{-1} \) where \( A^+ \) (respectively \( A^- \)) only gathers the positive (respectively negative) eigenvalues.}
Block ILU strategies for high order discretization of the Maxwell equations

boundary term called numerical flux (see also [11]). We refer to [7] for all the details
on the various terms involved in this DG formulation. Within each mesh element $K$
the electromagnetic field $(\mathbf{E}, \mathbf{H})^T$ is approximated as:

$$
(\mathbf{E}_h)_{|K} = \sum_{i=1}^{d_K} \mathbf{E}_K^i \varphi_i^K \quad \text{and} \quad (\mathbf{H}_h)_{|K} = \sum_{i=1}^{d_K} \mathbf{H}_K^i \varphi_i^K
$$

where $\mathbf{E}_K^i$ and $\mathbf{H}_K^i$ are the vectors of local degrees of freedom corresponding to the
basis expansion $\{\varphi_i^K\}_{i=1, \ldots, d_K}$ of $\mathbb{P}_p(K)$. In the present study, we adopt the classical
Lagrange nodal basis functions defined on a simplex and we assume that the inter-
polation degree is uniform (i.e. the same for all the elements of the mesh). Then the
resulting method is denoted as DG-$\mathbb{P}_p$.

4 Block preconditioning

The discretization of the system of time-harmonic Maxwell equations (3) using the
DG formulation (4) leads to a large sparse complex linear system of equations of
the form $\mathcal{A} \mathbf{W}_h \equiv (i\omega \mathcal{M} + \mathcal{C}) \mathbf{W}_h = \mathbf{b}$, where $\omega \mathcal{M}$ refers to the discretization of
the term:

$$
\int_{\Omega_h} (\omega G_0 \mathbf{W}_h)^T \nabla dV
$$

in (4), while $\mathcal{C}$ represents the discretization of the curl operators and the boundary
conditions for the remaining integrals on the left hand side of (4). For the numerical
treatment we assume that the sign of the first equation of the time-harmonic Maxwell
equations is flipped to $-i\omega \varepsilon_r \mathbf{E} + \text{curl} \mathbf{H} = \beta E \omega \varepsilon_r \mathbf{E} + \mathbf{J}$ and consistently changed
in $G_0, G_\beta, G_x, G_y, G_z$. In this case the matrices $\mathcal{M}$ and $\mathcal{C}$ become symmetric, thus
$\mathcal{A}$ is complex symmetric.

The matrix of this system exhibits a block structure which is linked to the use
of a polynomial interpolation method for the approximation of the electromagnetic
field within a mesh element (5). Up to a permutation which is induced by first taking
the contributions with respect to $\mathbf{E}$ and then the $\mathbf{H}$ part we find that:

$$
\mathcal{M} = \begin{pmatrix}
-M_{E_e} & 0 \\
0 & M_{\mu}\n\end{pmatrix}, \quad \mathcal{C} = \begin{pmatrix}
-C_{kk} & C_{kk}^T \\
C_{kk} & C_{kn} & C_{nn}\n\end{pmatrix},
$$

where $M_{E_e}$ and $M_{\mu}$ are real symmetric positive definite block diagonal matrices
whose block elements are the local mass matrices computed in each element $K$.
Computing a preconditioner based on an incomplete factorization of $\mathcal{A}$ happens to be prohibitively expensive. Therefore we shift the initial system by:

$$
\omega \left( -\beta_n M_{E_e} \quad 0 \\
0 & \beta_n M_{\mu}\right),
$$
where $\beta_k, \beta_h$ are chosen appropriately. This precisely refers to adding artificially $-\beta_k \omega \epsilon, E$ and $-\beta_h \omega \mu, H$ to the right-hand side of (1). With respect to $E$ this can be interpreted as artificial conductivity. We propose three different variants of block preconditioning. The first version consists of choosing $\beta_k = \beta_h = \beta$ and applying our preconditioner to the shifted system:

$$\mathcal{P}_1 = \beta \omega M + \mathcal{A}.$$ 

The second and third variant are best understood as a discrete analogy of eliminating the magnetic field $H$ from the second equation of the perturbed form of (1) and inserting it into the first equation of (1). The resulting equation thus reduces to:

$$\frac{1}{\omega (i + \beta_h)} \left( -(1 - \beta_k) (1 - \beta_h i) \omega \epsilon E + \text{curl} \left( \frac{1}{\mu} \text{curl} E \right) \right) = -J.$$ 

This is essentially a vector-valued Helmholtz equation, where the operator is shifted by a multiple of the mass matrix. The discrete analogy can be described by eliminating the $H$ part from $\beta \omega M + \mathcal{A}$ by one block elimination step:

$$\mathcal{S} = -\omega (i + \beta_h) M_{ee} - C_{eH} C_{H}^{T} \omega (i + \beta_h) M_{\mu} + C_{mH}^{-1} C_{eH}.$$ 

For the second variant block preconditioning we use $\beta = \beta_k = \beta_h$ to obtain the reduced system $\mathcal{P}_2$. This can be read as first shifting and then eliminating. Finally for the third variant we proceed analogously to the second one except that we first eliminate $H$ from the unshifted system $\mathcal{A}$ and then shift the reduced system by $-\beta \omega M_{\mu}$, i.e., we choose $\beta = \beta_k$ and $\beta_h = 0$ in order to obtain the reduced system $\mathcal{P}_3$. According to the work by Magolu [12], Erlangga et al [10], shifting the operator with a real-valued $\beta$ significantly improves incomplete LU preconditioning and multilevel preconditioning.

For preconditioning we apply the inverse-based multilevel block ILU [3], as implemented in ILUPACK$^2$. Its hallmark is the strategy of keeping the inverse triangular factors below a given bound $\kappa$. In order to deal with indefinite systems, a block factorization approach is used based on a symmetrized maximum weight matching (see [2] for details of this approach when being applied to Helmholtz equation).

5 Numerical results

We now present the impact of shifting the initial system by a multiple of the mass matrix for a 3D problem discretized by a $DG-P_1$ method. The problem under consideration is the scattering of a plane wave by a perfectly conducting unit sphere. The

---

$^2$ http://ilupack.tu-bs.de
frequency of the incident plane wave of frequency $F=900$ MHz and thus, we have $\omega = 18.84$ (after renormalization of the Maxwell equations). The computational domain is defined as the free space between the perfectly conducting sphere and an outer sphere on which the Silver-Müller absorbing condition is applied. We have used an unstructured tetrahedral mesh consisting of 46,704 tetrahedral elements. This yields a complex symmetric system of size $n = 1,120,896$. The computations were performed on a workstation equipped with an Intel Xeon E7440 CPU with frequency 2.4 GHz and 64 GB of memory. For the ILU we use a drop tolerance of $10^{-2}$ but limit the maximum amount of fill per row by $10 \times$ the number of nonzeros per row in $\mathcal{A}$. We use an inverse bound of $\kappa = 5$ for inverse-based pivoting. As iterative solver we use the simplified QMR method which allows for the use of (complex) symmetric systems and preconditioners. The iteration is stopped, whenever the backward error satisfies $\|Ax - b\| \leq 10^{-6}(\|A\| \|x\| + \|b\|)$. As comparison we also add numerical results of the direct solver PARDISO$^3$ (see Tab. 4). The numerical results in Tab. 1-3 confirm the efficiency of our shifted multilevel block ILU approach. They illustrate that shifting the initial system is essential for the ILU. If the shift is too small then the fill would increase drastically if there were no limit imposed. On the other hand, shifting the system too much turns the preconditioned system away from the original system.

**Table 1** Multilevel block ILU applied to $\mathcal{P}_1 = A + \beta \omega M$

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>ILU[sec]</th>
<th>$\frac{n c(ILU)}{n c(A)}$</th>
<th>levels</th>
<th>SQMR[sec]</th>
<th>steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>$8.8 \cdot 10^2$</td>
<td>11.7</td>
<td>5</td>
<td>$3.2 \cdot 10^3$</td>
<td>620</td>
</tr>
<tr>
<td>3.0</td>
<td>$1.7 \cdot 10^2$</td>
<td>5.4</td>
<td>4</td>
<td>$1.7 \cdot 10^3$</td>
<td>387</td>
</tr>
<tr>
<td>5.0</td>
<td>$1.0 \cdot 10^2$</td>
<td>6.2</td>
<td>2</td>
<td>$2.4 \cdot 10^3$</td>
<td>574</td>
</tr>
<tr>
<td>10.0</td>
<td>$4.6 \cdot 10^1$</td>
<td>3.3</td>
<td>1</td>
<td>$1.9 \cdot 10^3$</td>
<td>1035</td>
</tr>
</tbody>
</table>

**Table 2** Multilevel block ILU for the reduced system $\mathcal{P}_2$ of $A + \beta \omega M$ after eliminating the $E$ part first.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>ILU[sec]</th>
<th>$\frac{n c(ILU)}{n c(A)}$</th>
<th>levels</th>
<th>SQMR[sec]</th>
<th>steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>$3.6 \cdot 10^2$</td>
<td>9.9</td>
<td>6</td>
<td>$1.7 \cdot 10^3$</td>
<td>398</td>
</tr>
<tr>
<td>3.0</td>
<td>$1.4 \cdot 10^2$</td>
<td>5.2</td>
<td>2</td>
<td>$9.8 \cdot 10^2$</td>
<td>302</td>
</tr>
<tr>
<td>5.0</td>
<td>$8.5 \cdot 10^1$</td>
<td>4.3</td>
<td>2</td>
<td>$1.9 \cdot 10^3$</td>
<td>613</td>
</tr>
<tr>
<td>10.0</td>
<td>$3.9 \cdot 10^1$</td>
<td>1.9</td>
<td>1</td>
<td>$1.0 \cdot 10^3$</td>
<td>842</td>
</tr>
</tbody>
</table>

$^3$ http://www.pardiso-project.org
Table 3  Multilevel block ILU for the reduced system $\mathcal{P}_3$ of $A$ after eliminating the $E$ part first and then shifting by $\beta \omega M_\mu$.  

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$\text{ILU[sec]}$</th>
<th>$\frac{n_c(\text{ILU})}{n_c(A)}$</th>
<th>levels</th>
<th>$\text{SQMR[sec]}$</th>
<th>steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>$4.9 \cdot 10^2$</td>
<td>15.2</td>
<td>8</td>
<td>$9.7 \cdot 10^3$</td>
<td>1773</td>
</tr>
<tr>
<td>3.0</td>
<td>$3.5 \cdot 10^2$</td>
<td>9.2</td>
<td>5</td>
<td>$1.9 \cdot 10^3$</td>
<td>452</td>
</tr>
<tr>
<td>5.0</td>
<td>$2.6 \cdot 10^2$</td>
<td>6.7</td>
<td>4</td>
<td>$1.3 \cdot 10^3$</td>
<td>337</td>
</tr>
<tr>
<td>10.0</td>
<td>$1.6 \cdot 10^2$</td>
<td>6.0</td>
<td>3</td>
<td>$1.2 \cdot 10^3$</td>
<td>325</td>
</tr>
</tbody>
</table>

Table 4  Direct solver PARDISO applied to $A$  

<table>
<thead>
<tr>
<th>computation time</th>
<th>$\frac{n_c(\text{LU})}{n_c(A)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$5.2 \cdot 10^3$</td>
<td>90.4</td>
</tr>
</tbody>
</table>

References  