Research Article | Vol. 26, No. 13 | 25 Jun 2018 | OPTICS EXPRESS 16925
Optics EXPRESS

Accelerating convergence of iterative solution of finite difference frequency domain problems via schur complement domain decomposition

NATHAN ZHAO,[1] SACHA VERWEIJ,[1] WONSEOK SHIN,[2] AND SHANHUI FAN[3]

[1]*Department of Applied Physics, Stanford University, Stanford, California 94305, USA*
[2]*Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[3]*Department of Electrical Engineering, Stanford University, Stanford, California 94305, USA*
[*]*shanhui@stanford.edu*

Abstract: We show that iterative solution of Maxwell's equations using the finite-difference frequency-domain method can be significantly accelerated by using a Schur complement domain decomposition method. We account for the improvement by analyzing the spectral properties of the linear systems resulting from the use of the domain decomposition method.

OCIS codes: (000.4430) Numerical approximation and analysis; (050.1755) Computational electromagnetic methods; (220.0220) Optical design and fabrication; (350.4238) Nanophotonics and photonic crystals.

References and links

1. C. M. Soukoulis and M. Wegener, "Past achievements and future challenges in the development of three-dimensional photonic metamaterials," Nature Photonics **5**(9), 523–530 (2011).
2. N. Meinzer, W.L. Barnes, and I.R. Hooper, "Plasmonic meta-atoms and metasurfaces," Nature Photonics **8**(8), 889–898 (2014).
3. Z. Yu, G. Veronis, Z. Wang, and S. Fan, "One-way electromagnetic waveguide formed at the interface between a plasmonic metal under a static magnetic field and a photonic crystal," Physical Review Letters, **100**, 023902 2008.
4. N. Segal, S. Keren-Zur, N. Hendler, and T. Ellenbogen, "Controlling light with metamaterial-based nonlinear photonic crystals," Nature Photonics **8**(9), 180–184 (2015).
5. W. Cai and V. Shalaev, *Optical Metamaterials* (Springer, 2010).
6. S. Wang, P. Wu, V. Su, Y. Lai, C. Chu, J. Chen, S. Lu, J. Chen, B. Xu, C. Kuan, T. Li, S. Zhu, and D. Tsai, "Broadbrand achromatic optical metasurface devices," Nature Communications **8**, 187 (2017).
7. N. Yu and F. Capasso, "Flat optics with designer metasurfaces," Nature Materials **13**, 139–150 (2014).
8. B. J. Bohn, M. Schnell, M. A. Kats, F. Aieta, R. Hillenbrand, and F. Capasso, "Near-Field Imaging of Phased Array Metasurfaces", Nano Letters **15**(6), 3851–3858 (2015).
9. J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade, *Photonic crystals: putting a new twist on light* (Cambridge University Press, 2008)
10. A. Mekis, J.C. Chen, I. Kurland, S. Fan, P.R. Villeneuve, and J.D. Johannopoulos, "High Transmission through Sharp Bends in Photonic Crystal Waveguides," Phys. Rev. Lett. **77**(18), 3787–3790 (1996).
11. S. Fan, P.R. Villeneuve, and J.D. Joannopoulos, "Channel Drop Tunneling through Localized States," Phys. Rev. Lett. **80**(5), 960–963 (1998).
12. S. Verweij, V. Liu, and S. Fan, "Accelerating simulation of ensembles of locally differing optical structures via a Schur complement domain decomposition," Opt. Lett. **39**(22), 6458–6461 (2014).
13. Y. Li and J.M. Jin, "A vector dual-primal finite element tearing and interconnecting method for solving 3D large-scale electromagnetic problems," IEEE Transactions on Antennas and Propagation **54**(10), 704–723 (2006).
14. Zhen Peng, Kheng-Hwee Lim and Jin-Fa Lee, "Non-conformal Domain Decomposition Methods for Solving Large Multi-scale Electromagnetic Scattering Problems," Proceedings of the IEEE, **101**(2), 298–319 (2013).
15. Zhen Peng, Ralf Hiptmair, Yang Shao and Brian MacKie-Mason, "Domain decomposition preconditioning for surface integral equations in solving challenging electromagnetic scattering problems," IEEE Transactions on Antennas and Propagation **64**(1), 210–223 (2016).
16. K. Zhang and J.M. Jin, "Parallel FETI-DP algorithm for efficient simulation of large-scale electrmagnetic problems," International Journal of Numerical Modelling: Electronic Networks, Devices and Fields **29**(5), 897–914 (2016).
17. P. Jaysaval, D. Datta, M. Sen, and A. Arnulf, "A Schur complement based fast 2D finite-difference multimodel modeling of acoustic wavefield in the frequency domain," in SEG Technical Program Expanded Abstracts 2017, (Society of Exploration Geophysicists, 2017).

#327139 https://doi.org/10.1364/OE.26.016925
Journal © 2018 Received 5 Apr 2018; accepted 20 May 2018; published 18 Jun 2018

Optics EXPRESS

18. J. Maryska, M. Rozloznik and M. Tuma, "Schur Complement Systems in the Mixed-Hybrid Finite Element Approximation of the Potential Fluid Flow Problem," SIAM Journal on Scientific Computing **22**(2), 704–723 (2000).
19. B. Alavikia, N. Soltani, and O. M. Ramahi, "Efficient 2-D Finite-Difference Frequency-Domain Method for Switching Noise Analysis in Multilayer Boards," IEEE Transactions on Components, Packaging and Manufacturing Technology **3**(5), 841–848 (2013).
20. O. G. Ernst, M.J. Gander, "Why it is difficult to solve Helmholtz problems with classical iterative methods," in Numerical Analysis of Multiscale Problems, I.G. Graham, T.Y. Hou, O. Lakkis, and R. Scheichl ed. (Springer Berlin Heidelberg, 2012).
21. W. Shin, "Choice of the perfectly matched layer boundary condition for frequency-domain Maxwell's equations solvers," Journal of Computational Physics **231**(8), 3406–3431 (2012).
22. K. S. Kunz, R.J. Luebbers, The Finite Difference Time Domain Method for Electromagnetics, CRC-Press, 1993. Section 3.2
23. V. Simoncini and D.B. Szyld, "Recent computational developments in Krylov subspace methods for linear systems," Numerical Linear Algebra with Applications **14**(1), 1–59 (2007).
24. W. Shin and S. Fan. "Simulation of phenomena characterized by partial differential equations," US Patent Application 13/744,999
25. J. Liesen and P. Tichy, "Convergence analysis of Krylov subspace methods," GAMM-Mitteilungen **27**(2), 153–173 (2004).
26. W. Shin and S. Fan, "Accelerated solution of the frequency-domain Maxwell's equations by engineering the eigenvalue distribution of the operator," Opt. Express **21**(19), 22578–22595 (2013).
27. J. Mandel, "On block diagonal and Schur complement preconditioning," Numerische Mathematik **58**(1), 79–93 (1990).
28. A. R. Horn. and F. Zhang, *The Schur complement and its applications* (Springer, 2005), Chap. 1.
29. A. Pyzara, B. Bylina and J. Bylina, "The influence of a matrix condition number on iterative methods convergence," in Federated Conference on Computer Science and Information Systems (FedCSIS, 2011), pp. 459–464.
30. M. Arioli and F. Romani, "Relations between condition numbers and the convergence of the Jacobi method for real positive definite matrices," Numerische Mathematik **46**(1), 31–42 (1985).
31. M. Neytcheva, "On element-by-element Schur complement approximations," Linear Algebra and its Applications **434**(11), 2308–2324 (2011).
32. M. Storti, L. Dalcin, R. Paz, A. Yommi, V. Sonzogni and N. Nigro, "A preconditioner for the Schur complement matrix," Advances in Engineering Software **37**(11), 754–762 (2006).
33. L. Kulas and M. Mrozowski, "Low-reflection subgridding," IEEE Transactions on Microwave Theory and Techniques **53**(5), 1587–1592 (2005).
34. T. Wu and Z. Chen, "A dispersion minimizing subgridding finite difference scheme for the Helmholtz equation with PML," Journal of Computational and Applied Mathematics **267**, 82–95 (2014).
35. L. Giraud, A. Haidar, and Y. Saad, "Sparse approximations of the Schur complement for parallel algebraic hybrid linear solvers in 3D," [Research Report] RR-7237, INRIA. 2010, pp.18.
36. Y. Saad and B. Suchomel, "ARMS: an algebraic recursive multilevel solver for general sparse linear systems," Numerical Linear Algebra with Applications **9**(5), 359–378 (2002).

1. Introduction

Nanophotonic structures often consist of large collections of repeated meta-atoms as in, for example, many functional photonic crystals or metamaterials [1–11]. Furthermore, the individual meta-atoms often contain fine, subwavelength features that makes simulating large collections of these meta-atoms challenging. In simulating these structures, finding a way to exploit the high degree of repeated structure could provide significant efficiency improvements.

A simple and natural way to exploit the repetition of meta-atoms in such structures is to use a non-overlapping domain decomposition approach, for example a Schur complement domain decomposition [12]. Domain decomposition has been investigated for nanophotonics extensively in the context of finite element methods using sophisticated iterative substructuring techniques, for example in finite element tearing and interconnect (FETI) methods [13, 16]. Additionally, variations focusing on solution of scattering problems using surface integral methods as in [14, 15] have also been studied. However, given the simplicity of finite difference methods relative to finite element methods, investigating simple domain decomposition methods for the former is worthwhile, and we will show that applying a Schur complement domain decomposition to domains consisting of many repeated meta-atoms proves quite advantageous.

In the domain decomposition approach, one solves Maxwell's equations on a large domain by

Optics EXPRESS

breaking that domain into smaller subdomains, and solving relatively small problems associated with each. In the Schur complement domain decomposition, one transforms the original problem on the large domain by first solving Maxwell's equations on a number of smaller subdomains in isolation. One then constructs a secondary, usually smaller, problem on the interfaces between subdomains to account for coupling between the subdomains. One uses the solution to this interfacial problem, combined with the solutions to the subdomain problems, to recover the solution on the complete domain. To take advantage of repeated meta-atoms, one can solve the subproblem associated with a repeated meta-atom once for all meta-atoms of the same kind, and reuse the solution at negligible cost wherever necessary.

The domain decomposition approach thus reduces the original problem to a problem with many fewer degrees of freedom, namely those associated with the interfaces between the subdomains. Such an interfacial problem is cheap to form and small in dimension relative to the original system. Also, the domain decomposition approach is useful in design optimization problems since one can solve an associated subproblem on subdomains that do not change during the optimization, and then perform calculations only on the subdomain involving the optimization region and the interfaces [12, 17–19].

Furthermore, iterative solution of Maxwell's equations using the standard finite-difference frequency-domain method can suffer from indefiniteness and poor conditioning of the matrix resulting from discretization, even when standard preconditioning methods are used [20]. We show that applying a Schur complement domain decomposition can mitigate such problems, leading to substantial improvements in convergence rate. Additionally, we show that the ratio of source wavelength to subdomain size is an important consideration in achieving such improvement.

The rest of this paper proceeds as follows. Section 2 describes the Schur complement domain decomposition via an example, using it to accelerate the convergence of iterative solution of a finite-difference frequency-domain (FDFD) discretization of Maxwell's equations. Section 3 provides several numerical experiments showing improvements in convergence when applying the Schur complement domain decomposition method. Section 4 shows how the numerical results in Sec. 3 can be heuristically understood by analyzing the conditioning of the matrices that arise from the domain decomposition method. Section 5 addresses memory and wall-clock time requirements of the domain decomposition method, and briefly comments on the use of subgridding to further enhance the method, particularly in three dimensions. Finally Sec. 6 provides some concluding remarks.

2. Formulation of the Schur complement domain decomposition

For most of this paper, we will consider two-dimensional structures such as the structure shown in Fig. 1. The structure consists of a collection of metallic or dielectric objects distributed on a square lattice with lattice constant a. The objects and the bounding cell will be referred to as meta-atoms. The space between the objects in each meta-atom is air. Either a periodic boundary condition or a perfectly matched layer (PML) boundary condition surrounds the domain.

Figure 1 shows an exemplary structure on an $8a \times 8a$ lattice. For illustration purposes we consider three types of objects: square (with a side length of $0.5a$), isosceles right triangle (with a base length of $0.33a$), and circle (with a radius of $0.25a$). Each such object resides at the center of a meta-atom. We assume that the objects all have the same dielectric constant ϵ. Throughout this paper, we consider two cases, the *dielectric case*, where $\epsilon = 12$ approximates the dielectric constant of Si and GaAs in the infrared wavelength range, and the *metallic case*, where $\epsilon = -3 - 0.3i$ is the complex dielectric constant for a metal with some loss. The choice of the metallic dielectric constant was made so that the resolution of the grid picked for the dielectric case would be sufficient to resolve the surface waves that would appear on the metal. The structure shown here does not have any apparent optical functionality, and serves only to demonstrate the algorithms. We note, however, that many published works on nanophotonic structures, including

Optics EXPRESS

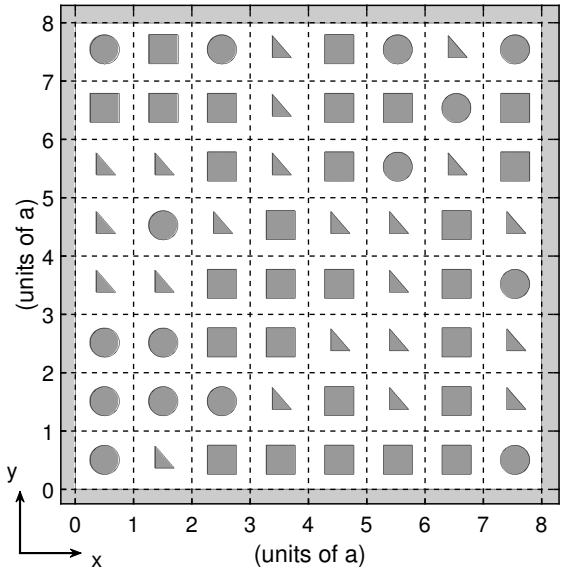

Fig. 1. Two-dimensional simulation domain consisting of an 8×8 grid of square meta-atoms. Each meta-atom contains a dielectric inclusion shown as a gray shape. A point source is placed in the center. The dotted lines denote the interfaces between meta-atoms. The lightly shaded region on the borders of the domain represents the perfectly matched layer boundary condition.

photonic crystals, meta-materials, and meta-surfaces, consider similar structures, where a set of dielectric or metallic objects (not necessarily identical) are distributed on a regular lattice [1–11].

We solve for the field distribution inside the domain induced by a point source at the center of the domain. For concreteness, we consider the TE polarization where the magnetic field **H** has only one component perpendicular to the plane [9], which we designate the z-component. The governing equation is then:

$$\left[\left(\frac{\partial}{\partial x}\epsilon^{-1}(x,y)\right)\frac{\partial}{\partial x} + \left(\frac{\partial}{\partial y}\epsilon^{-1}(x,y)\right)\frac{\partial}{\partial y}\right]H_z + \omega^2\mu_0 H_z = i\omega M_z \tag{1}$$

where H_z denotes the perpendicular component of the magnetic field, μ_0 the magnetic permeability, M_z the magnetic current source, ω the angular frequency of the source, and $\epsilon(x, y)$ the dielectric distribution of the domain. Throughout this paper we choose an angular frequency corresponding to a free space wavelength of $1.5a$.

A standard way to solve Eq. (1) is the finite-difference frequency-domain method (FDFD). In this method, one discretizes the H_z field on a regular grid. In our simulation the grid spacing is $0.012a$. Approximating Eq. (1) with finite differences on that grid yields a linear system of the form $Ax = b$. This linear system, when describing a domain consisting of a collection of meta-atoms like that of Fig. 1, can be written:

$$\begin{bmatrix} A_{00} & B_{01} & B_{02} & \cdots & B_{0n} \\ C_{10} & D_{11} & 0 & \cdots & 0 \\ C_{20} & 0 & D_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ C_{n0} & 0 & 0 & 0 & D_{nn} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \tag{2}$$

The vector x_0 corresponds to the field components that lie at the interfaces between different meta-atoms, as indicated by the dashed lines in Fig. 1. x_i for $i \neq 0$ corresponds to the field

Optics EXPRESS

components in the interior of the ith meta-atom. Each diagonal block in A corresponds to a restriction of the approximation to Eq. (1) to a subdomain and terminated with a Dirichlet boundary condition. The block A_{00} describes the coupling of field components within the union of the boundaries, or interfaces, of all the meta-atoms, which we call the interface. Each block D_{ii} describes the couplings between field components within the interior of each meta-atom. The blocks B_{0i} and C_{i0} describes the couplings between the field components at the boundary of the interior of the ith meta-atom with the field components at the interface of the ith meta-atom. Finally, b_i contains information about the sources within the ith meta-atom.

To solve Eq. (2) via the domain decomposition approach, we solve for each x_i in terms of x_0:

$$x_i = D_{ii}^{-1}(b_i - C_{i0}x_0) \tag{3}$$

Then, we substitute every such expression for x_i into the equation obtained from the first row of A in Eq. (2), yielding a reduced system expressed only in terms of x_0:

$$Sx_0 = \left(A_{00} - \textstyle\sum_{i=1}^{N} B_{0i}D_{ii}^{-1}C_{i0}\right)x_0 = b_0 - \displaystyle\sum_{i=1}^{N} B_{0i}D_{ii}^{-1}b_i \tag{4}$$

S is the Schur complement of the block diagonal matrix $D = diag(D_{11}, D_{22}, ..., D_{nn})$ in A. b' will denote the modified source term on the right hand side of Eq. (4). This linear system contains only the degrees of freedom on the interfaces shown in Fig. 1. In this paper, we refer to the system in Eq. (4) as the *reduced system*, and the system in Eq. (2) as the *unreduced system*. There is a natural correspondence between solving these two systems whenever D is not singular since one can recover the full solution on the total domain after solving Eq. (4) by substituting the solution x_0 into Eq. (3). Numerically, however, the reduced system S is much smaller in dimension than A, often up to an order of magnitude or more, meaning it could be much more efficient to solve Eq. (4) in practice.

To explicitly form S in Eq. (4), one must evaluate the action of D_{ii}^{-1} on the C_{0i} blocks, a potentially computationally expensive step. But where many of the meta-atoms are identical, and consequently many of the sub matrices D_{ii} are identical, explicitly forming S requires only evaluating a suitable factorization (to calculate the action of the inverse of D_{ii}) of every distinct D_{ii} once and reusing them at negligible or much lower expense to calculate $D_{ii}^{-1}C_{i0}b_i$ and $B_{0i}D_{ii}^{-1}C_{i0}$. For general systems with repeated meta-atoms, the interface degrees of freedom in x_0 should be selected to minimize the number of distinct subdomains, and hence minimize the number of distinct factorizations of D_{ii}. In addition, there is a memory cost since S is a denser matrix than A due to fill-in from $B_{0i}D_{ii}^{-1}C_{i0}$. In spite of these upfront or additional costs, we will show that in many cases, solving Eq. (4) can be significantly better than solving Eq. (2). Additionally, one can bypass the costs mentioned above of explicitly forming S by instead repeatedly applying the action of S to a vector during iterative solution. Such a method is not atypical of Krylov subspace iterative solvers [23].

3. Main numerical results

3.1. Methods

In this section, we demonstrate the advantage of solving Eq. (4) over Eq. (2) in simulating the system described in Sec. 2. To solve these linear systems, we apply an iterative method, which is preferable over direct methods for large systems [23]. Here we use the method of quasi-minimal residual (QMR), which has been demonstrated to work with the FDFD method, particularly for systems with inhomogenous dielectric distributions [21]. We also have tried a few other iterative solvers such as the generalized minimal residual (GMRES) and biconjugate gradient (BICG) method, neither of which converge as fast as QMR for our systems (but produce qualitatively similar results).

Optics EXPRESS

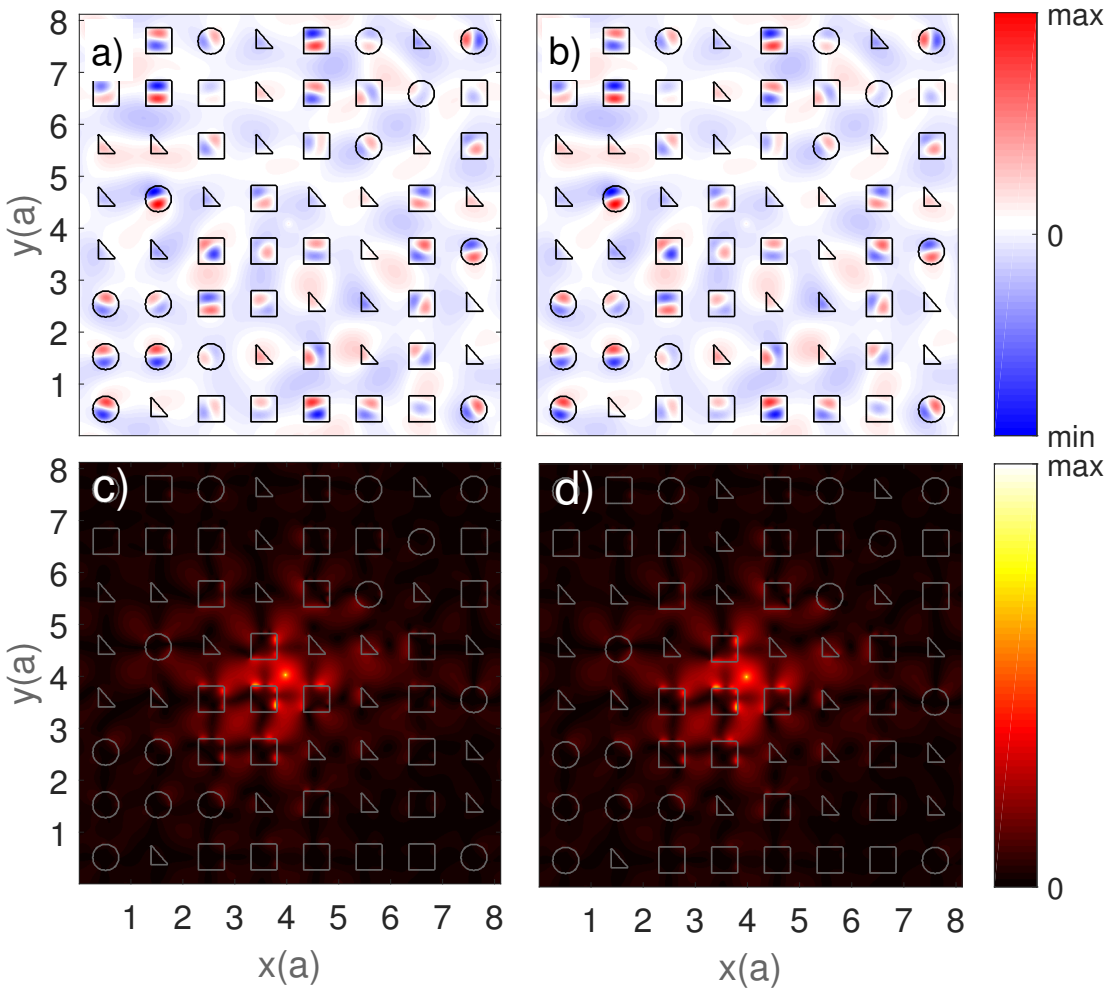

Fig. 2. The first column shows the fields (real for dielectric, absolute value for metallic) from the iterative solution (using QMR) of Eq. (4) for a) the dielectric case ($\epsilon = 12$) and c) the metallic case ($\epsilon = -3 - 0.3i$). The second column shows the equivalents obtained with a direct solver for b) the dielectric case and d) the metallic case. Since the fields are complex in the metallic case, we show the absolute value of the fields. The source wavelength in all cases is $1.5a$

3.2. *Results for two-dimensional systems*

First, we verify that the solution recovered from the reduced system using Eq. (3) is accurate. To that end, since our test case is small enough, we use a direct solver to robustly and accurately produce a solution from the unreduced system (Eq. (2)) for comparison. In Fig. 2 we consider the structure in Fig. 1 surrounded by a periodic boundary condition. Figures 2(a) and 2(b) correspond to the dielectric case, where both the matrix A in Eq. (2) and the matrix S in Eq. (4) are real symmetric. Figures 2(c) and 2(d) correspond to the metallic case where both A and S are complex symmetric. In both cases, we see that the solutions from the direct solver (Figs. 2(a) and 2(c)) agree well with the solutions obtained from the domain decomposition approach with the solution to x_0 acquired with an iterative solver (Figs. 2(b) and 2(d)). The relative norm error between these two solutions in both cases is less than 10^{-10}.

To compare the convergence rates of iteratively solving Eq. (2) versus Eq. (4), we plot the

Optics EXPRESS

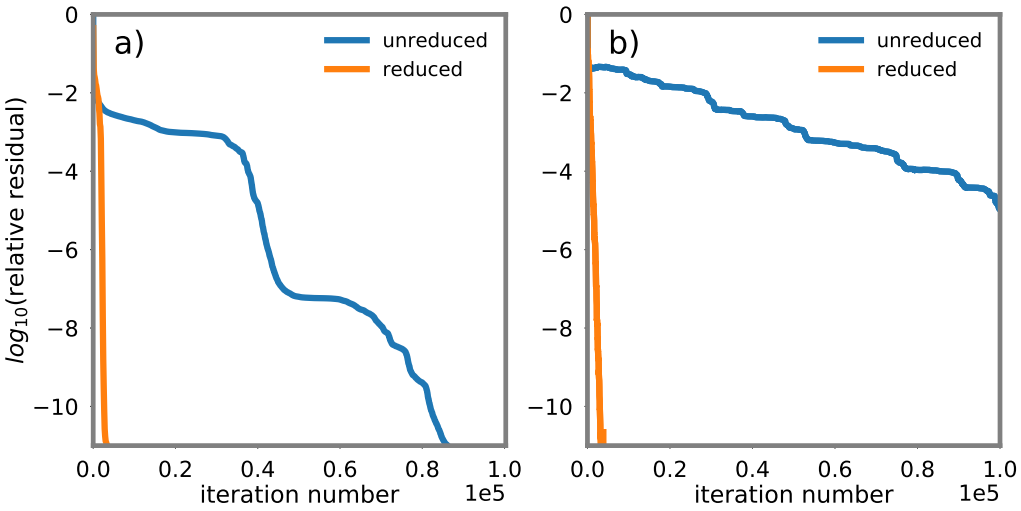

Fig. 3. Convergence rates for the unreduced and reduced systems on the domain containing an 8×8 lattice of meta-atoms from Fig. 1 with a periodic boundary condition for a) the dielectric case and b) the metallic case. The source wavelength in both cases is $1.5a$.

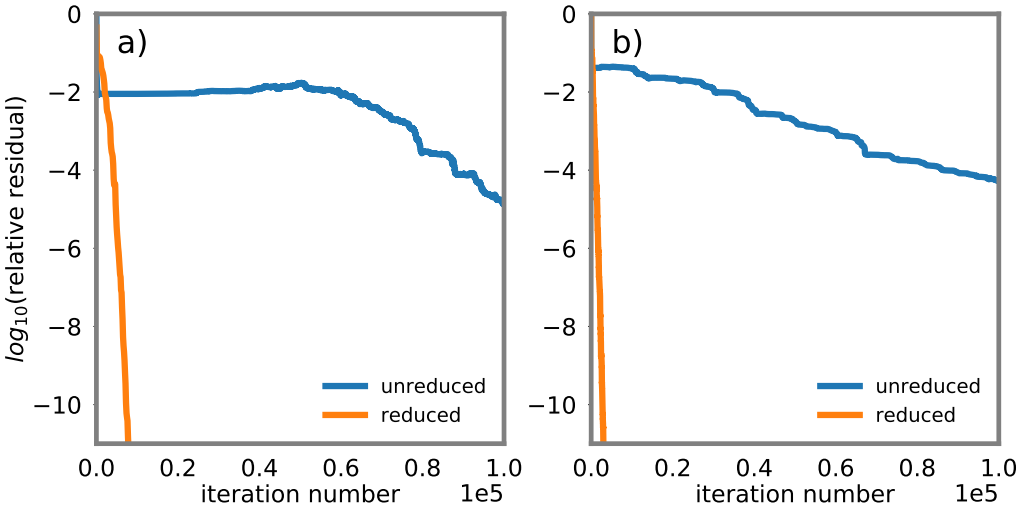

Fig. 4. Comparison of convergence rates of the unreduced and reduced systems on the domain containing an 8×8 lattice of meta-atoms from Fig. 1 with a perfectly matched layer 15 grid points deep in a) the dielectric case and b) the metallic case. The source wavelength in both cases is $1.5a$

Optics EXPRESS

relative residual error $\rho = ||Ax - b||/||b||$ for a solution x as a function of the number of iterations. For the unreduced system, x is obtained directly by solving $Ax = b$. For the reduced system, we solve for x_0 in Eq. (4) and then obtain x using Eq. (3).

Figure 3 shows the relative residual error as a function of the number of iterations. For both the dielectric case (Fig. 3(a)) and the metallic case (Fig. 3(b)), we see that the number of iterations to reach a relative residual error $|\rho| < 10^{-11}$ is roughly an order of magnitude lower when solving the reduced system than when solving the unreduced system. We note however that there is a scale difference between the dielectric and metallic cases. The dielectric case in general appears to require significantly more iterations to solve than a comparable metallic system.

Figure 4 shows the convergence rates of the same two systems considered in Fig. 3, but with a stretched-coordinate perfectly-matched-layer (PML) boundary condition surrounding the domain. The perfectly matched layer is 15 grid points deep. When applying the domain decomposition method, the perfectly matched layer is split into 36 subdomains as shown in Fig. 1. For both the dielectric system (Fig. 4(a)) and the metallic system (Fig. 4(b)), the unreduced system exhibits some stagnation while the reduced system still converges. Even with the preconditioner from [24], which symmetrizes the PML-terminated FDFD linear system, the problem persists. For both cases considered in Figs. 3 and 4, the convergence improvements from solving the reduced system also translate to significant wall-clock time improvements, an observation further discussed in Sec. 5.

In Fig. 5 we focus on the dielectric case, and plot the number of iterations to convergence as a function of domain size. For each domain consisting of an $N \times N$ lattice of meta-atoms, we randomly place one of the three inclusions discussed above (circle, triangle, square) in each meta-atom. We place a point source at the center of the domain with a free-space wavelength of $1.5a$. In Figs. 5(a) and 5(b) we consider the cases with either a perfectly-matched-layer or periodic boundary condition respectively, and compare the performance of iteratively solving the reduced and the unreduced cases. In all cases, the number of iterations to convergence increases as the domain expands. For the systems with a periodic boundary condition, the factor of improvement, defined as the ratio between the number of iterations to convergence for the unreduced and reduced systems, remains approximately constant (Fig. 5(a)). For the systems with a perfectly matched layer boundary condition, the unreduced system fails to converge when the domain size exceeds $8a$, whereas the reduced system converges for all domain sizes considered, as shown in Fig. 5(b). We also note that in the simulations with a PML, simply applying the domain decomposition exclusively to the PML region suffices to make the iterative solver converge. In general, we observe that applying the domain decomposition method significantly improves convergence rates for the finite-difference frequency-domain discretizations of Maxwell's equations discussed here.

3.3. Extension to three-dimensional systems

The domain decomposition method discussed above can be applied in three dimensions as well. In three dimensions, we solve:

$$\nabla\times\nabla\times\mathbf{E} - \nabla\left[\frac{1}{\epsilon(x,y,z)}\nabla\cdot\left(\epsilon(x,y,z)\mathbf{E}\right)\right] - \omega^2\mu_0\epsilon(x,y,z)\mathbf{E} = -i\omega\mu_0\mathbf{J} - \frac{i}{\omega}\nabla\left[\frac{1}{\epsilon(x,y,z)}\nabla\cdot\mathbf{J}\right] \quad (5)$$

where $\mathbf{E}$ is the electric field and $\mathbf{J}$ the current source. Eq. (5) is a formulation of Maxwell's equations; the gradient terms in Eq. (5), absent in most formulations of Maxwell's equations, make the resulting matrix better conditioned [26].

To illustrate the domain decomposition method, we consider the structure shown in Fig. 6. The structure is a $3 \times 3 \times 3$ lattice of cubic meta-atoms containing dielectric or metallic cubes of side length $a/2$ where a is the lattice constant. We use the same dielectric constants for the cubes as in the dielectric and metallic cases for the two-dimensional simulations. A dipole source with

Optics EXPRESS

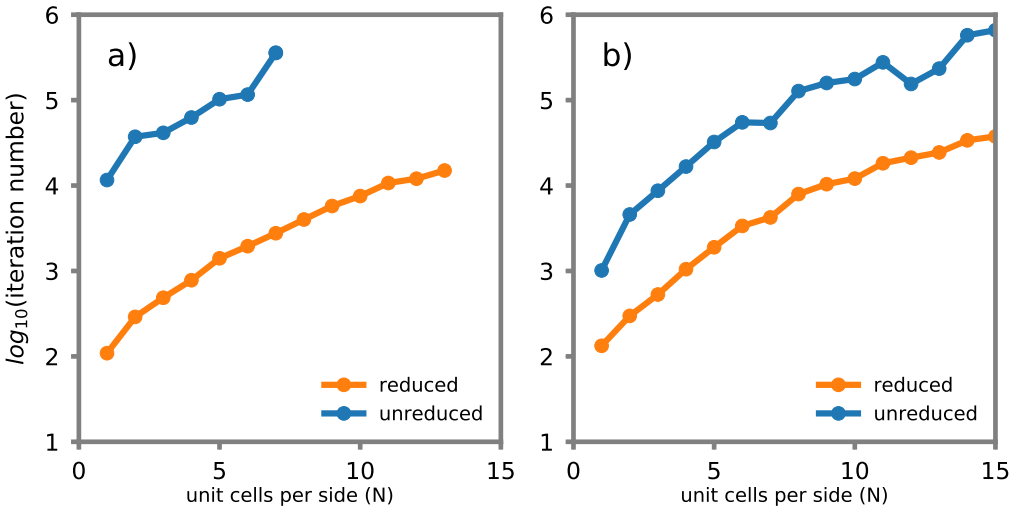

Fig. 5. Number of iterations to convergence for the unreduced and reduced systems, as a function of increasing domain size. Here, the horizontal axis shows the number of meta-atoms per side in an $N \times N$ lattice of meta-atoms. We consider the dielectric case with a) perfectly matched layer boundary condition and b) periodic boundary condition. The source wavelength is $1.5a$.

a free-space wavelength of $2.5a$ resides at the interface between the central meta-atom and a diagonal neighbor meta-atom (so the source is not inside the dielectric/metallic inclusion inside the central meta-atom). We discretize Eq. (5) with a Yee's grid with a grid spacing of $0.053a$. A periodic boundary condition terminates the domain.

For this structure, we consider the convergence rate of solving the reduced and unreduced systems. In the unreduced case, we directly solve Eq. (5) using the method discussed in [26]. In the reduced case, we solve Eq. (4); the x_0 in Eq. (3) corresponds to the field components at the boundaries between neighboring meta-atoms. In the dielectric case (Fig. 7(a)), iterative solution of the unreduced system stagnates while that of the reduced system converges. In the metallic case (Fig. 7(b)), solving the reduced system converges significantly faster than solving the unreduced system.

Unlike the two-dimensional systems considered in the previous section, in three dimensions, for example the metallic case above, the reduction in the number of iterations to convergence does not necessarily translate to a reduction in net computational cost since the matrix for the reduced system can be quite dense. A number of ways to circumvent this problem exist; Sec. 5 provides a more detailed discussion of this issue.

4. Heuristics for understanding convergence improvements

The previous section provided numerical observations that the reduced system typically converges faster than the unreduced system. This section presents intuition regarding such observations and discusses the specific conditions under which the reduced system shows improvement over the unreduced system.

4.1. Mathematical background

As a basis for our heuristic argument, there is an extensive mathematics literature regarding the convergence behavior of iterative solution of symmetric positive definite linear systems $Ax = b$.

Optics EXPRESS

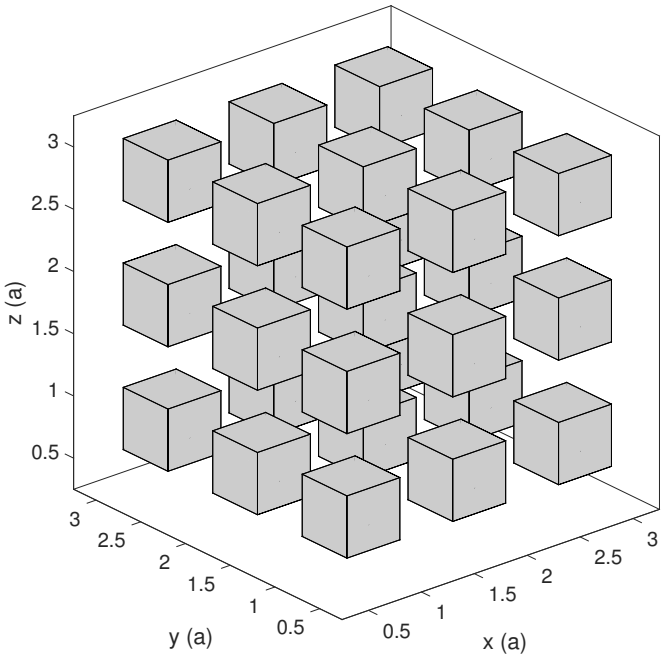

Fig. 6. A lattice of cubic meta-atoms containing dielectric or metallic cubes. The volume between the dielectric or metallic cubes is air.

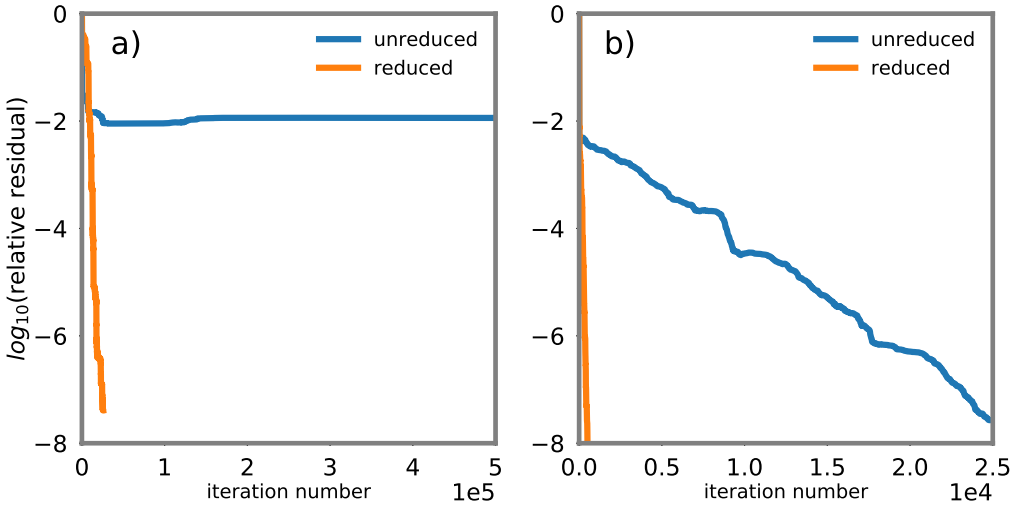

Fig. 7. Comparison of convergence rates for the unreduced and reduced systems terminated by a periodic boundary condition. a) corresponds to the dielectric case and b) the metallic case. In both cases, the reduced system converges in fewer iterations than the unreduced system. The source wavelength is 2.5a.

Optics EXPRESS

For such systems, some insight into convergence behavior can be gained from analyzing the condition number of the matrix A:

$$\mathrm{cond}(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \tag{6}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are A's maximum and minimum singular values.

The condition number bounds the convergence rate of specific classes of iterative solvers applied to symmetric positive definite linear systems [29, 30]. Though in general one cannot use condition number to understand the convergence behavior of linear systems that are not symmetric positive definite, in practice it often proves a useful guide. In that spirit, here we compare the conditioning of the unreduced system A with the reduced system S.

The Schur complement S of any sub-block D in a positive definite matrix A is at least as well conditioned as the original matrix A [27, 28]. This weak bound can be further strengthened under various conditions. For a matrix A in the form of Eq. (2), [27, 31] show that:

$$\mathrm{cond}(S) \leq \frac{\mathrm{cond}(A_{00})}{1 - \gamma_B^2} \tag{7}$$

In Eq. (7) γ_B^2 is the Cauchy-Bunyakowski-Schwarz (CBS) constant, given by:

$$\gamma_B^2 = \rho(A_{00}^{-1} B D^{-1} C) \tag{8}$$

where $\rho(X)$ is the spectral radius of matrix X (the absolute value of the largest magnitude eigenvalue of the matrix). The CBS constant measures the relative weight of the off-diagonal blocks to the diagonal blocks. According to [31], for any symmetric positive semi-definite matrix, the CBS constant is bounded above by one (equal to one if the matrix is singular). Moreover, the value of γ_B^2 is close to zero when the norm of the diagonal blocks is much larger than the norm of the off-diagonal blocks. In general,for a symmetric positive definite matrix A, the minimum requirement for S to be better conditioned than A is that γ_B^2 is not close to 1 and $\mathrm{cond}(A_{00}) < \mathrm{cond}(A)$.

4.2. Application to FDFD

The results above apply only to symmetric positive definite matrices. For other matrices, particularly indefinite matrices, the applicability of the results in the previous section is not guaranteed. However, in this section, we prove that an extreme conditioning contrast between A_{00} and A is possible in FDFD systems, which matches the theoretical requirement that $\mathrm{cond}(A_{00}) < \mathrm{cond}(A)$ discussed above. Additionally, we know that appropriate FDFD formulations of Maxwell's equations, such as those used above, can result in systems that are very close to being spectrally definite [26], which may make the arguments in the previous section more applicable in practice. We also note that, in numerical experiments like those above, not using the appropriate formulations of Maxwell's equations as mentioned above often diminished the reduced system's convergence improvement over the unreduced system's, or even made the reduced system's convergence poorer than the unreduced system's.

In the domain decomposition approach above, the diagonal block A_{00} in Eq. (8) corresponds to the degrees of freedom in the domain in Fig. 1 on the interfaces of the meta-atoms. Below, we will show that A_{00} is far better conditioned than the FDFD matrix A. Therefore, Eq. (8) indicates that the reduced matrix S can be better conditioned than A, which we show is numerically true for our systems even though they are indefinite.

For the matrix A of the unreduced system, its singular values ($\sigma_i(A)$) can be estimated as:

$$\sigma_i(A) = \left| -\omega_{i,A}^2 + \omega^2 \right| \mu_0 \tag{9}$$

Optics EXPRESS

where $\omega_{i,A}$ denotes the ith eigenfrequency of the derivative operator in Eq. (1) and ω denotes the source frequency. The maximum and minimum values of $\omega_{i,A}$ do not necessarily determine $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$, but can serve as lower and upper bounds respectively. The largest magnitude eigenfrequency of the derivative operator on a regular grid can be approximated $\omega_{\max,A} \approx (c_A\pi/\Delta)$, where Δ is the finite-difference grid spacing [21, 22]. c_A is the speed of light in the medium in the subdomain with the smallest index of refraction. In typical simulations the source frequency is far smaller than $\omega_{\max,A}$. Therefore, A's maximum singular value is:

$$\sigma_{\max,A} = \left| -\omega_{\max,A}^2 + \omega^2 \right| \mu_0 \approx \left| -\left(c_A \frac{\pi}{\Delta}\right)^2 \right| \mu_0 \tag{10}$$

A_{00}'s maximum singular value is also approximately determined by Eq. (10), since A_{00} is the matrix A restricted to the interfacial degrees of freedom, and the grid spacing remains the same on this subdomain. As a result, $\sigma_{max}(A_{00}) \approx \sigma_{\max}(A)$.

To compare the smallest singular values of A and A_{00}, we first note that the smallest eigenfrequency $\omega_{min,A}$ of the derivative operator is overall determined by the size of the full domain as $\omega_{min,A} \approx \sqrt{\sum_{j=1}^{d} (c_A\pi/L_j)^2}$, where L_js are side lengths of the full domain and j indexes the spatial dimensions (d) to sum over. Also, the driving wavelength is typically chosen comparable to the lattice constant a in order to maximize the interaction between the wave and the meta-atoms. We have $a \ll L_j$ because the simulation domain is composed of several meta-atoms in each direction. Therefore, we have

$$\sigma_{\min}(A) = \min_i \left| -\omega_{i,A}^2 + \omega^2 \right| \mu_0 \leq \left| -\omega_{\min,A}^2 + \omega^2 \right| \mu_0 \approx \omega^2 \mu_0 \tag{11}$$

Finally, we can approximate $\omega^2\mu_0$ since the source wavelength is approximately the size of a:

$$\omega^2 \mu_0 \sim (c_A\pi/a)^2 \mu_0 \tag{12}$$

On the other hand, the interface is very thin, typically only one or a few grid spacings thick, and therefore its thickness is much less than the transverse dimensions of the interface. Therefore, we have $\omega_{\min,A_{00}} = \sqrt{\sum_j (c\pi/L_{j,interface})^2} \approx c_{A_{00}}\pi/t$, where t is the thickness of the interface. Since t is much less than the lattice constant a, we have

$$\sigma_{\min}(A_{00}) = \min_i \left| -\omega_{i,A_{00}}^2 + \omega^2 \right| \mu_0 \leq \left| -\omega_{\min,A_{00}}^2 + \omega^2 \right| \mu_0 \approx \left(c_{A_{00}} \frac{\pi}{t} \right)^2 \mu_0. \tag{13}$$

Combining the results so far, we conclude

$$\sigma_{\min}(A_{00}) \approx (c_{A_{00}}\pi/t)^2 \mu_0 \gg (c_A\pi/a)^2 \mu_0 \gtrsim \sigma_{\min}(A). \tag{14}$$

Combining the arguments in the preceding paragraphs for both the maximum and minimum singular values, we therefore have $\mathrm{cond}(A_{00}) \ll \mathrm{cond}(A)$. Now that we have proved that for FDFD, we can achieve large conditioning contrast between A_{00} and A, we now demonstrate numerically that $\mathrm{cond}(S) \ll \mathrm{cond}(A)$. In Fig. 8, we consider the same system as in Fig. 5 for the dielectric case. We plot the condition numbers of A_{00}, S, and A as a function of domain size, with the domain surrounded by either a periodic boundary condition (Fig. 8a), or a perfectly matched layer boundary condition (Fig. 8b). In both Figs. 8a and 8b, the condition number of A_{00} is largely independent of domain size, and is four orders of magnitude or more below the condition number of A. The condition number of S is around two orders of magnitude lower than the condition number of A. Such a large contrast in conditioning between A_{00} and A, combined with the experimental observation that $\mathrm{cond}(S) \ll \mathrm{cond}(A)$, suggests that we can apply some of

Optics EXPRESS

the insights from Sec. 4.1 to our system. In particular, the size of the conditioning improvement between A and A_{00} is important because, were A symmetric positive definite, this means that conditioning improvement of S is still possible even if $\gamma_B^2 \approx 1$, as long as $1/|1-\gamma_B^2|$ is less than the magnitude of the ratio of the conditioning of A over A_{00}.

This conditioning analysis provides an intuitive explanation for why the use of domain decomposition techniques can significantly improve convergence in FDFD, as observed in Sec. 4. This analysis also indicates that in applying this domain decomposition technique, if possible it is better to choose the interfaces between subdomains to be in the region with lower dielectric constant, since in this case the condition number of A_{00} should be smaller due to a higher value of $\sigma_{min}(A_{00})$.

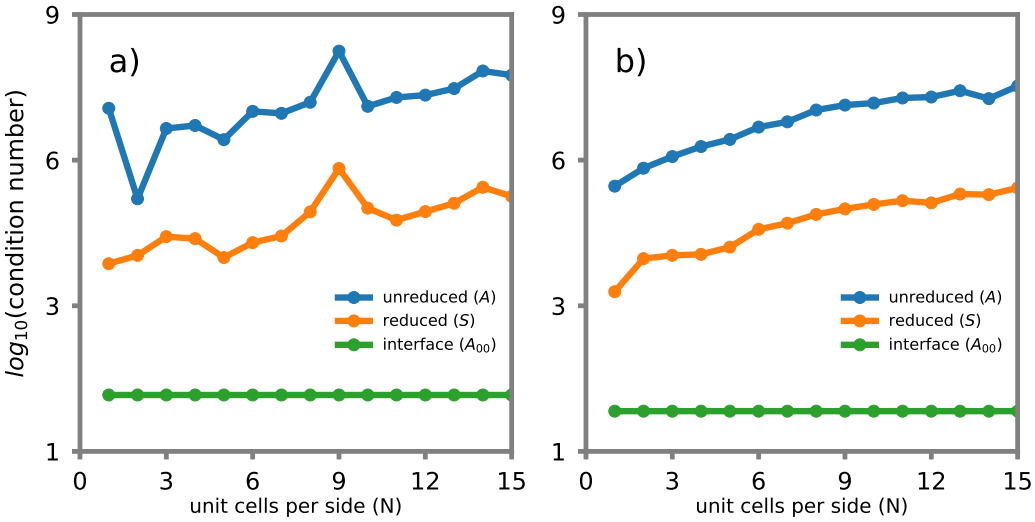

Fig. 8. Condition number of the unreduced matrix (A), the reduced matrix (S), and the interface sub-block of A (A_{00}), as a function of domain size, for the physical system described in Fig. 5 where the dielectric inclusions have $\epsilon = 12$ with (a) periodic and (b) perfectly matched layer boundary conditions.

5. Memory and wall-clock time considerations

The time required for iterative solution of a linear system $Ax = b$ is roughly $T_A I_A$, where T_A denotes the time to evaluate the matrix-vector product Ax, and I_A represents the number of iterations to convergence. The preceding analysis focused on I_A. This section discusses T_A in the situation where we explicitly form S, and highlights directions for future work, especially in three dimensions.

When solving a linear system iteratively, the number of nonzero elements in the matrix roughly determines both the memory necessary to store the system, as well as the time for each matrix-vector product (matvec). To simplify our discussion, we consider a domain consisting of a d-dimensional cubic lattice of cubic meta-atoms. Additionally, we consider the thinnest possible interface we can construct. For example, with $d = 2$, such an interface consists of a layer one grid-spacing thick between each meta-atom, as illustrated by the dotted lines in Fig. 1. We define P_t as the number of nodes per side length of each meta-atom and N the number of meta-atoms along each dimension (for a total of N^d meta-atoms in the grid for a spatial dimension d). $C(d)$ denotes the number of E-field components where d is the spatial dimension of the system. When $d = 3$, $C(3) = 3$ (see Eq. (5)) and when $d = 2$, $C(2) = 1$ (see Eq. (1)). Finally, $\Sigma(d)$ denotes the

Optics EXPRESS

size of the stencil for each node, which takes on a value of $\Sigma(3) = 13$ in three dimensions for the standard Yee's grid and $\Sigma(2) = 5$ for a typical compact stencil in two dimensions. Hence, the unreduced system A contains approximately $\Sigma(d)C(d)(NP_t)^d$ nonzero entries. Then for A, the time cost for a matrix-vector product scales approximately as:

$$T_A \propto \Sigma(d)C(d)[NP_t]^d = \begin{cases} 5[NP_t]^2, & d = 2 \\ 39[NP_t]^3, & d = 3 \end{cases} \tag{15}$$

To discuss the reduced system, we let P_e denote the number of nodes per side length of the interface surrounding each meta-atom. In general $P_e \neq P_t$ unless the grid spacing on the interface is the same as that on the interior. The total number of nonzero elements in the reduced matrix S is then approximately $N^d(dC(d)P_e^{d-1})^2$. To get this expression, we observe that for each of the N^d meta-atoms, we get a dense sub-block of the form $B_{0i}D_{ii}^{-1}C_{i0}$. We count the number of degrees of freedom on the interface of each sub-block, which is $dC(d)P_e^{d-1}$. Since the block is fully dense, we must then square this term. We also note that for $d = 3$, $C(3) = 3$ is not quite accurate as the thinnest possible interface on the Yee grid that decouples two adjacent meta-atoms only requires two components. However, to maintain consistent notation, we use $C(3) = 3$ as a strong upper bound. Finally, the same cost for S is roughly:

$$T_S \propto N^d[dC(d)P_e^{d-1}]^2 = \begin{cases} N^2[2P_e]^2, & d = 2 \\ N^3[9P_e^2]^2, & d = 3 \end{cases} \tag{16}$$

Now taking the ratio of Eq. (15) over Eq. (16):

$$\frac{T_A}{T_S} \approx \frac{\Sigma(d)C(d)P_t^d}{\left(dC(d)P_e^{(d-1)}\right)^2} = \begin{cases} \frac{5P_t^2}{4P_e^2}, & d = 2 \\ \frac{13P_t^3}{27P_e^4}, & d = 3 \end{cases} \tag{17}$$

First, consider a uniform grid such that $P_e = P_t$. In two dimensions, evaluating Eq. (17) gives $T_A/T_S \approx 5/4$, independent of P_t or P_e. A ratio less than one indicates how much more costly evaluating a matvec for S is versus A. Conversely, a ratio greater than one means matvecs with S are cheaper to evaluate than A. Therefore, given a factor of 10 improvement in the convergence rates as demonstrated in Sec. 3, we see that in the two-dimensional case, significant wall-clock speed-up is possible. In three dimensions, however, the ratio in Eq. (17) acquires a factor of $1/P_t$. Thus, both the memory cost for storing the matrix and the time required for matrix-vector products makes solving the reduced system less attractive than solving the unreduced system unless the improvement in I_A is substantially better than the increase in T_A. In the three-dimensional case considered in Sec. 5, for example, where $P_t = 19$, the storage for the reduced system is roughly a factor of 40 greater than that for the unreduced system. However, one can simply evaluate the action of S as discussed in Sec. 2, rather than forming and applying S explicitly, to bypass this issue.

For three-dimensional systems, in order to achieve significant reduction in computational cost via the domain decomposition approach described above, one can consider subgridding [33, 34]. Specifically, one might discretize the interfaces between meta-atoms more coarsely than the interiors of meta-atoms ($\alpha P_e = P_t$ in Eq. (13), where α is a positive integer greater than 1). Doing so could reduce the memory cost and matrix-vector product evaluation time for the reduced system relative to the unreduced system. Additionally, subgridding is natural for the types of systems considered in this paper. The need to capture all complex geometry, as well as all relatively rapid field variation within large-dielectric-constant materials, forces a fine discretization scale. In physical systems like these, all such complex geometry and large-dielectric-constant material typically appears within the meta-atoms' interiors rather than on their interfaces. Thus, in principle the interfaces between meta-atoms need not be discretized at the same resolution as the interiors of meta-atoms.

Optics EXPRESS

6. Conclusions and final remarks

In this paper, we consider iterative solution of finite-difference frequency-domain discretizations of Maxwell's equations, and show that a Schur complement domain decomposition method can significantly improve convergence rates when considering commonly used nanophotonic structures consisting of many repeated meta-atoms. The improvement in convergence rates can be heuristically understood in terms of the improvement in condition number of the relevant matrices.

The success of the approach here exploits the fact that the size of the meta-atoms is comparable to or smaller than the free space wavelength. The ability of the domain decomposition method to reuse solutions on subdomains containing repeated meta-atoms should facilitate optimization of photonic structures based on repeated meta-atoms. Further improvement of the method can be achieved via subgridding, as discussed in Sec. 5.

Aside from subgridding, one could precondition the reduced system S during iterative solution [32], which is important as the reduced system's iterations to convergence and condition number increases with increasing domain size as exhibited in Fig. 5 and Fig. 8 respectively. Additionally, work has been done to find sparse approximations of the Schur complement [35], which can also mitigate storage costs mentioned in Sec. 5. Moreover, there are many preconditioning schemes on the unreduced linear system Eq. (2) that exploit domain decomposition and also avoid directly solving or iterating with S. Particular examples include algebraic recursive multi-level schemes (ARMS), which are described in [36]. The development here points to the possibility of using these methods to accelerate electromagnetic simulations.

Funding

Air Force Office of Scientific Research (AFOSR) (FA9550-15-1-0335, FAFA9550-17-1-0002); Defense Advanced Research Project Agency (HR00111720034); U.S. Department of Energy (DOE) Office of Science (SC) (DE-AC05-06OR23100).

Acknowledgments

The authors thank Jerry (Yu) Shi for illuminating discussions about the subtleties of the finite difference frequency domain method.