Accelerated solution of the frequency-domain Maxwell's equations by engineering the eigenvalue distribution of the operator

Wonseok Shin[1,2] **and Shanhui Fan**[1,*]

[1]*Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA*
[2]*wsshin@stanford.edu*
[*]*shanhui@stanford.edu*

Abstract: We introduce a simple method to accelerate the convergence of iterative solvers of the frequency-domain Maxwell's equations for deep-subwavelength structures. Using the continuity equation, the method eliminates the high multiplicity of near-zero eigenvalues of the operator while leaving the operator nearly positive-definite. The impact of the modified eigenvalue distribution on the accelerated convergence is explained by visualizing residual vectors and residual polynomials.

OCIS codes: (000.4430) Numerical approximation and analysis; (000.3860) Mathematical methods in physics; (000.3870) Mathematics.

References and links

1. N. J. Champagne II, J. G. Berryman, and H. M. Buettner, "FDFD: A 3D finite-difference frequency-domain code for electromagnetic induction tomography," J. Comput. Phys. **170**, 830–848 (2001).
2. G. Veronis and S. Fan, "Overview of simulation techniques for plasmonic devices," in "Surface Plasmon Nanophotonics," , M. Brongersma and P. Kik, eds. (Springer, 2007), pp. 169–182.
3. U. S. Inan and R. A. Marshall, *Numerical Electromagnetics: The FDTD Method* (Cambridge University, 2011). Ch. 14.
4. J.-M. Jin, *The Finite Element Method in Electromagnetics*, 2nd ed. (Wiley, 2002).
5. V. Simoncini and D. B. Szyld, "Recent computational developments in Krylov subspace methods for linear systems," Numer. Linear Algebra Appl. **14**, 1–59 (2007).
6. W. Cai, W. Shin, S. Fan, and M. L. Brongersma, "Elements for plasmonic nanocircuits with three-dimensional slot waveguides," Adv. Mater. **22**, 5120–5124 (2010).
7. G. Veronis and S. Fan, "Bends and splitters in metal-dielectric-metal subwavelength plasmonic waveguides," Appl. Phys. Lett. **87**, 131102–3 (2005).
8. L. Verslegers, P. Catrysse, Z. Yu, W. Shin, Z. Ruan, and S. Fan, "Phase front design with metallic pillar arrays," Opt. Lett. **35**, 844–846 (2010).
9. J. T. Smith, "Conservative modeling of 3-D electromagnetic fields, Part II: Biconjugate gradient solution and an accelerator," Geophys. **61**, 1319–1324 (1996).
10. G. A. Newman and D. L. Alumbaugh, "Three-dimensional induction logging problems, Part 2: A finite-difference solution," Geophys. **67**, 484–491 (2002).
11. C. J. Weiss and G. A. Newman, "Electromagnetic induction in a generalized 3D anisotropic earth, Part 2: The LIN preconditioner," Geophys. **68**, 922–930 (2003).
12. V. L. Druskin, L. A. Knizhnerman, and P. Lee, "New spectral Lanczos decomposition method for induction modeling in arbitrary 3-D geometry," Geophys. **64**, 701–706 (1999).
13. E. Haber, U. M. Ascher, D. A. Aruliah, and D. W. Oldenburg, "Fast simulation of 3D electromagnetic problems using potentials," J. Comput. Phys. **163**, 150–171 (2000).
14. J. Hou, R. Mallan, and C. Torres-Verdin, "Finite-difference simulation of borehole EM measurements in 3D anisotropic media using coupled scalar-vector potentials," Geophys. **71**, G225–G233 (2006).

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

15. R. Hiptmair, F. Kramer, and J. Ostrowski, “A robust Maxwell formulation for all frequencies,” IEEE Trans. Magn. **44**, 682–685 (2008).
16. K. Beilenhoff, W. Heinrich, and H. Hartnagel, “Improved finite-difference formulation in frequency domain for three-dimensional scattering problems,” IEEE Trans. Microwave Theory Tech. **40**, 540–546 (1992).
17. A. Christ and H. L. Hartnagel, “Three-dimensional finite-difference method for the analysis of microwave-device embedding,” IEEE Trans. Microwave Theory Tech. **35**, 688–696 (1987).
18. At the final stage of our work, we were made aware of a related work by M. Kordy, E. Cherkaev, and P. Wannamaker, “Schelkunoff potential for electromagnetic field: proof of existence and uniqueness” (to be published), where an equation similar to our Eq. (7) with $s = -1$ was developed.
19. A. Jennings, “Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method,” IMA J. Appl. Math. **20**, 61–72 (1977).
20. A. van der Sluis and H. van der Vorst, “The rate of convergence of conjugate gradients,” Numer. Math. **48**, 543–560 (1986).
21. S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer, “GMRES and the minimal polynomial,” BIT Numer. Math. **36**, 664–675 (1996).
22. S. Goossens and D. Roose, “Ritz and harmonic Ritz values and the convergence of FOM and GMRES,” Numer. Linear Algebra Appl. **6**, 281–293 (1999).
23. M. Benzi, G. H. Golub, and J. Liesen, “Numerical solution of saddle point problems,” Acta Numer. **14**, 1–137 (2005). Sec. 9.2.
24. B. Beckermann and A. B. J. Kuijlaars, “Superlinear convergence of conjugate gradients,” SIAM J. Numer. Anal. **39**, 300–329 (2001).
25. B. Beckermann and A. B. J. Kuijlaars, “On the sharpness of an asymptotic error estimate for conjugate gradients,” BIT Numer. Math. **41**, 856–867 (2001).
26. O. Axelsson, “Iteration number for the conjugate gradient method,” Math. Comput. Simulat.**61**, 421–435 (2003).
27. J. P. Webb, “The finite-element method for finding modes of dielectric-loaded cavities,” IEEE Trans. Microwave Theory Tech. **33**, 635–639 (1985).
28. F. Kikuchi, “Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism,” Comput. Method Appl. Mech. Eng. **64**, 509–521 (1987).
29. D. A. White and J. M. Koning, “Computing solenoidal eigenmodes of the vector Helmholtz equation: a novel approach,” IEEE Trans. Magn. **38**, 3420–3425 (2002).
30. N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, 1st ed. (Saunders College, 1976), Ch. 8.
31. To obtain $8/\Delta_{\text{min}}^2$, take the first equation of Eq. (4.29) in [32] and then multiply the extra factor $\mu = \mu_0$ to account for the difference between Eq. (4.17a) in [32] and Eq. (1) in the present paper.
32. W. Shin and S. Fan, “Choice of the perfectly matched layer boundary condition for frequency-domain Maxwell’s equations solvers,” J. Comput. Phys. **231**, 3406–3431 (2012).
33. Y. Saad and M. H. Schultz, “GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems,” SIAM J. Sci. Stat. Comp. **7**, 856–869 (1986).
34. P. B. Johnson and R. W. Christy, “Optical constants of the noble metals,” Phys. Rev. B **6**, 4370–4379 (1972).
35. E. D. Palik, ed., *Handbook of Optical Constants of Solids* (Academic Press, 1985).
36. D. R. Lide, ed., *CRC Handbook of Chemistry and Physics*, 88th ed. (CRC, 2007).
37. R. Freund and N. Nachtigal, “QMR: a quasi-minimal residual method for non-hermitian linear systems,” Numer. Math. **60**, 315–339 (1991).
38. R. W. Freund, G. H. Golub, and N. M. Nachtigal, “Iterative solution of linear systems,” Acta Numer. **1**, 57–100 (1992). Secs. 2.4 and 3.3.
39. J. W. Goodman, *Introduction to Fourier Optics* (Roberts & Company Publishers, 2005), 3rd ed. Sec. 2.3.2.

1. Introduction

To understand electromagnetic (EM) and optical phenomena, it is essential to solve Maxwell’s equations efficiently. In the frequency domain, assuming a time dependence $e^{+i\omega t}$ and nonmagnetic materials, Maxwell’s equations reduce to

$$\nabla \times \nabla \times \mathbf{E} - \omega^2 \mu_0 \varepsilon \mathbf{E} = -i\omega\mu_0 \mathbf{J}, \tag{1}$$

where ε is the electric permittivity (which can be complex); μ_0 is the magnetic permeability of vacuum; ω is the angular frequency; **E** and **J** are the electric field and electric current source density, respectively.

To solve Eq. (1) numerically, one can use a method such as the finite-difference frequency-domain (FDFD) method [1–3] or the finite element method (FEM) [4] to construct a large

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

system of linear equations

$$Ax = b, \tag{2}$$

where A is a matrix representing the operator $[\nabla \times (\nabla \times \quad) - \omega^2 \mu_0 \varepsilon]$; x is an unknown column vector representing $\mathbf{E}$; b is a column vector representing $-i\omega\mu_0\mathbf{J}$. The matrix A thus constructed is sparse (with only 13 nonzero elements per row when generated by the FDFD method) and typically very large (often with more than 10 million rows and columns for three-dimensional (3D) problems). To solve a system with such a large and sparse matrix, iterative methods are usually preferred to direct methods [5].

However, in the "low-frequency regime" where the wavelength is much longer than the grid cell size Δ, it is well-known that convergence is quite slow when the iterative methods are directly applied to solve Eq. (1). The low-frequency regime arises, for example, in nanophotonics [6–8] and geophysics [9–11] where structures have feature sizes that are at deep-subwavelength scale, and it will be defined more rigorously in Sec. 2. The huge null space of the operator $\nabla \times (\nabla \times \quad)$ was shown to be the origin of the slow convergence [10,11], and several techniques to improve the convergence speed have been developed.

The first class of techniques is based on the Helmholtz decomposition, which decomposes the E-field as $\mathbf{E} = \boldsymbol{\Psi} + \nabla\varphi$, where $\boldsymbol{\Psi}$ is a divergence-free vector field and φ is a scalar field [9–15]. Because $\nabla \cdot \boldsymbol{\Psi} = 0$, Eq. (1) is written as

$$-\nabla^2 \boldsymbol{\Psi} - \omega^2 \mu_0 \varepsilon (\boldsymbol{\Psi} + \nabla\varphi) = -i\omega\mu_0 \mathbf{J}, \tag{3}$$

where the operator $\nabla \times (\nabla \times \quad)$, which has a huge null space, is replaced with the negative Laplacian $-\nabla^2$, which is positive-definite for appropriate boundary conditions and thus has the smallest possible null space. However, these techniques either solve an extra equation for the extra unknown φ at every iteration step [9–12], which can be time-consuming, or increase the number of the rows and columns of the matrix by about 33% [13–15], which requires more memory.

The second class of techniques utilizes the charge-free condition

$$\nabla \cdot (\varepsilon \mathbf{E}) = 0. \tag{4}$$

The condition (4) holds at every source-free (i.e., $\mathbf{J} = 0$) position, where Eq. (1) can be modified to

$$\nabla \times \nabla \times \mathbf{E} + s\nabla [\nabla \cdot ((\varepsilon/\varepsilon_0)\mathbf{E})] - \omega^2 \mu_0 \varepsilon \mathbf{E} = 0 \tag{5}$$

for an arbitrary constant s; note that the right-hand side is 0 because $\mathbf{J} = 0$. In this class of techniques, Eqs. (1) and (5) are solved at positions with and without sources, respectively.

Reference [16] applied the above technique with $s = +1$ to boundary value problems described in [17] and achieved accelerated convergence. Such boundary value problems satisfied $\mathbf{J} = 0$ everywhere, so Eq. (5) was solved throughout the entire simulation domain.

However, Ref. [16] did not conduct a detailed comparison of convergence speed between different values of s. It also did not report whether its technique leads to accelerated convergence for problems with sources, even though many problems have nonzero electric current sources $\mathbf{J}$ inside the simulation domain. Reference [1] applied the technique with $s = +1$ to problems with sources, but only in order to suppress spurious modes rather than to accelerate convergence.

In this paper, we develop a modification of Eq. (1) that improves convergence speed even if electric current sources $\mathbf{J}$ exist inside the simulation domain [18]. Unlike the previous technique that made the modification only at source-free positions, our technique modifies Eq. (1) everywhere including positions with sources. For the modification, we utilize the continuity equation

$$i\omega\rho + \nabla \cdot \mathbf{J} = 0, \quad \text{or} \quad \nabla \cdot (\varepsilon \mathbf{E}) = \frac{i}{\omega} \nabla \cdot \mathbf{J}, \tag{6}$$

which can be derived by taking the divergence of Eq. (1). When Eq. (6) is manipulated appropriately and then added to Eq. (1), we obtain

$$\nabla \times \nabla \times \mathbf{E} + s\nabla\left[\varepsilon^{-1}\nabla \cdot (\varepsilon \mathbf{E})\right] - \omega^2 \mu_0 \varepsilon \mathbf{E} = -i\omega\mu_0 \mathbf{J} + s\frac{i}{\omega}\nabla\left[\varepsilon^{-1}\nabla \cdot \mathbf{J}\right] \tag{7}$$

for a constant s. The modified equation (7) is the equation to solve in this paper.

The solution E-field of Eq. (7) is the same as the solution of the original equation (1) regardless of the value of s, because the solution of Eq. (1) always satisfies Eq. (6). However, the choice of s affects the convergence speed of iterative methods significantly. In this paper, we demonstrate that $s = -1$ induces faster convergence speed than other values of s by comparing the convergence behavior of iterative methods for $s = -1, 0, +1$; the latter two values of s are of particular interest, because $s = 0$ reduces Eq. (7) to the original equation (1) and $s = +1$ is the value that Ref. [16] used in Eq. (5), which is similar to Eq. (7).

We also show that the difference in convergence behavior results from the different eigenvalue distributions of the operators for different s. There are many general mathematical studies about the dependence of the convergence behavior on the eigenvalue distribution [19–26]. Our aim here is instead to provide an intuitive understanding of the convergence behavior specifically for the operator of Eq. (7). For this purpose, at each iteration step we visualize the residual vector and residual polynomial, which are widely used concepts to explain the convergence behavior of iterative methods [5] and also defined briefly in Sec. 3. As a result, we find that convergence speed deteriorates substantially for $s = 0$ because the operator has eigenvalues clustered near zero, and for $s = +1$ because the operator is strongly indefinite.

The rest of this paper is organized as follows. In Sec. 2 we investigate the eigenvalue distribution of the operator in Eq. (7) for $s = 0, -1, +1$ for a simple homogeneous system. We also define the low-frequency regime rigorously in the section. In Sec. 3, we relate the eigenvalue distribution with the convergence behavior of an iterative method. In Sec. 4, we solve Eq. (7) for a wide range of realistic 3D problems to compare the convergence behavior of an iterative method for the three values of s, and we conclude in Sec. 5.

2. Eigenvalue distribution of the operator for a homogeneous system

In this section, we consider the operator in Eq. (7) for a homogeneous system and show that the properties of the eigenvalue distribution of the operator strongly depend on the value of s. The impact of s on the eigenvalue distribution has been studied in detail in the literature of the deflation method (also known as the penalty method) [27–29]. Here we only highlight those aspects that are important for the present study.

For a homogeneous system where ε is constant, Eq. (7) is simplified to

$$\nabla \times \nabla \times \mathbf{E} + s\nabla(\nabla \cdot \mathbf{E}) - \omega^2 \mu_0 \varepsilon \mathbf{E} = -i\omega\mu_0 \mathbf{J} + s\frac{i}{\omega\varepsilon}\nabla(\nabla \cdot \mathbf{J}), \tag{8}$$

where the operator

$$T = \nabla \times (\nabla \times \quad) + s\nabla(\nabla \cdot \quad) - \omega^2 \mu_0 \varepsilon \tag{9}$$

is Hermitian for real ε. Because ε is constant in this section, the eigenvalue distribution of T is shifted from the eigenvalue distribution of a Hermitian operator

$$T_0 = \nabla \times (\nabla \times \quad) + s\nabla(\nabla \cdot \quad) \tag{10}$$

by a constant $-\omega^2 \mu_0 \varepsilon$. In the low-frequency regime such shift is negligible, and thus the eigenvalue distribution of T_0 approximates that of T very well. Hence, we examine the eigenvalue distribution of T_0 below to investigate the eigenvalue distribution of T.

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

Table 1. Properties of the eigenvalue distributions of T_0 for different s. Depending on the sign of s, T_0 has very different eigenvalue distributions in terms of the multiplicity of the eigenvalue 0 and the definiteness of T_0.

	$s = 0$	$s < 0$	$s > 0$
multiplicity of $\lambda = 0$	very high	low	
definiteness of T_0	positive-semidefinite		indefinite

In Appendix A, we show that $\mathbf{F}_\mathbf{k} e^{-i\mathbf{k}\cdot\mathbf{r}}$ with

$$\mathbf{F}_\mathbf{k} = \begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix}, \begin{bmatrix} k_z \\ 0 \\ -k_x \end{bmatrix}, \begin{bmatrix} -k_y \\ k_x \\ 0 \end{bmatrix} \tag{11}$$

are the three eigenfunctions of both $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$ for each wavevector $\mathbf{k}$. We also show in the same appendix that the corresponding three eigenvalues are

$$\lambda = 0,\ |\mathbf{k}|^2,\ |\mathbf{k}|^2 \tag{12}$$

for $\nabla \times (\nabla \times \quad)$, and

$$\lambda = -|\mathbf{k}|^2,\ 0,\ 0 \tag{13}$$

for $\nabla(\nabla \cdot \quad)$. Therefore, T_0 has

$$\lambda = -s|\mathbf{k}|^2,\ |\mathbf{k}|^2,\ |\mathbf{k}|^2 \tag{14}$$

as three eigenvalues for each wavevector $\mathbf{k}$.

Equation (14) indicates that the eigenvalue distribution of T_0 is greatly affected by the value of s. Specifically, the multiplicity of the eigenvalue 0 depends critically on whether s is 0 or not: for $s = 0$ T_0 has a very high multiplicity of the eigenvalue 0 because Eq. (14) has 0 as an eigenvalue for every $\mathbf{k}$, whereas for $s \neq 0$ T_0 does not have such a high multiplicity of the eigenvalue 0. The definiteness of T_0 also depends on the value of s: for $s \leq 0$ T_0 is positive-semidefinite because the three eigenvalues in Eq. (14) are always nonnegative, whereas for $s > 0$ T_0 is indefinite because Eq. (14) has both positive and negative numbers as eigenvalues. The different properties of the eigenvalue distributions of T_0 for $s = 0$, $s < 0$, and $s > 0$ are summarized in Table 1.

The above description of the eigenvalue distributions of T_0 should approximately hold for the eigenvalue distributions of T as well in the low-frequency regime, as mentioned in the discussion of Eqs. (9) and (10). Moreover, even though T is a differential operator defined in an infinite space, it turns out that the description also applies to the matrix A discretized from T that is defined in a spatially bounded simulation domain.

To demonstrate, we numerically calculate the eigenvalues of A for a two-dimensional (2D) system shown in Fig. 1, a square domain filled with vacuum. The domain is discretized on a finite-difference grid with $N_x \times N_y = 50 \times 50$ cells and cell size $\Delta = 2\,\mathrm{nm}$. Therefore, the matrix A for each s has $3N_xN_y = 7500$ rows and columns, where the extra factor 3 accounts for the three Cartesian components of the E-field. We choose ω corresponding to the vacuum wavelength $\lambda_0 = 1550\,\mathrm{nm}$, which puts the system in the low-frequency regime as will be seen at the end of this section. The matrices A are constructed for three values of s: 0, -1, and $+1$, each of which represents each category of s in Table 1.

The distributions of the numerically calculated eigenvalues of A for $s = 0, -1, +1$ are shown as three plots in Fig. 2. In each plot, the horizontal axis represents eigenvalues, and it is divided

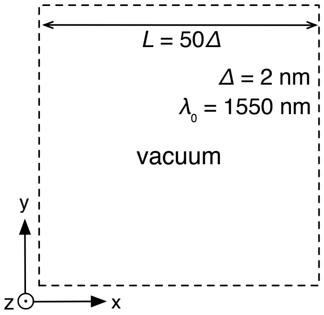
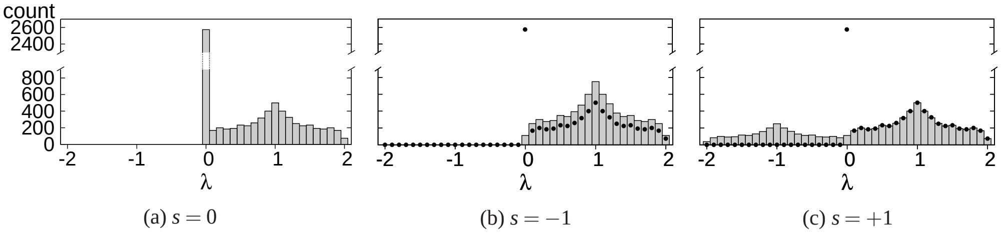


Fig. 1. A 2D square domain filled with vacuum ($\varepsilon = \varepsilon_0$) for which the eigenvalue distribution of T is calculated numerically for $s = 0, -1, +1$. The domain is homogeneous in the z-direction, whereas its x- and y-boundaries are subject to periodic boundary conditions. The square domain is discretized on a finite-difference grid with cell size $\Delta = 2\,\text{nm}$. The domain is composed of 50×50 grid cells, which lead to 7500 eigenvalues in total. A vacuum wavelength $\lambda_0 = 1550\,\text{nm}$, which puts the system in the low-frequency regime, is assumed for the electric current source to be used in Sec. 3.



(a) $s = 0$

(b) $s = -1$

(c) $s = +1$

Fig. 2. The eigenvalue distribution of A discretized from T for (a) $s = 0$, (b) $s = -1$, and (c) $s = +1$ for the vacuum-filled domain illustrated in Fig. 1. All 7500 eigenvalues λ of A are calculated for each s and categorized into 41 intervals in the horizontal axis that represents the range of the eigenvalues; the unit of the horizontal axis is nm^{-2}. The height of the column on each interval represents the number of the eigenvalues in the interval. In (b) and (c), the black dots indicate the eigenvalue distribution for $s = 0$ shown in (a). The vertical axes are broken due to the extremely tall column at $\lambda \simeq 0$ in (a). The local maxima at $\lambda = \pm 1\,\text{nm}^{-2}$ are the Van Hove singularities [30] arising from the lattice structure imposed by the finite-difference grid.

into 41 intervals $t_{-20}, \ldots, t_0, \ldots, t_{20}$ where $t_0 \ni 0$. The height of the column on each interval corresponds to the number of the eigenvalues in the interval.

The eigenvalue distributions of A shown in Fig. 2 agree well with the description of the eigenvalues of T_0 in Table 1: the very tall column on t_0 in Fig. 2(a) indicates the very high multiplicity of $\lambda \simeq 0$ for $s = 0$, and the eigenvalues distributed over $t_{j<0}$ and $t_{j>0}$ in Fig. 2(c) indicate a strongly indefinite operator for $s = +1$. In addition, the height of the column on t_0 in Fig. 2(a) is about 2500, or one third of the total number of eigenvalues, which agrees with Eq. (14) for $s = 0$ where one of the three eigenvalues is 0 for each $\mathbf{k}$; the columns on $t_{j>0}$ are about 1.5 times taller in Fig. 2(b) than in Fig. 2(a), which also agrees with Eq. (14) where the number of $|\mathbf{k}|^2$ increases from two for $s = 0$ to three for $s = -1$.

We end the section by providing a quantitative definition of the low-frequency regime. Suppose that A_0 is the matrix discretized from T_0 of Eq. (10). For $s = 0$, the eigenvalues of A_0 range

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

from 0 to $8/\Delta_{\min}^2$, where $\Delta_{\min}$ is the minimum grid cell size [31]; note that the range agrees with Fig. 2(a). The eigenvalue distribution of A is the shifted eigenvalue distribution of A_0 by $-\omega^2\mu_0\varepsilon$. The low-frequency regime is where the magnitude of the shift is so small that A has an almost identical eigenvalue distribution as A_0. Therefore, the condition for the low-frequency regime is

$$\omega^2\mu_0|\varepsilon| \ll 8/\Delta_{\min}^2. \tag{15}$$

Equation (15) is consistent with the condition introduced in [10], but here we provide a condition that is based on a more accurate estimate of the maximum eigenvalue of A_0. We can rewrite Eq. (15) in terms of the vacuum wavelength λ_0 as

$$\lambda_0/\Delta_{\min} \gg \pi\sqrt{|\varepsilon_r|/2}, \tag{16}$$

where $\varepsilon_r = \varepsilon/\varepsilon_0$ is the relative electric permittivity. The system described in Fig. 1 satisfies Eq. (16), so it is in the low-frequency regime.

3. Impact of the eigenvalue distribution on the convergence behavior of GMRES

In this section, we explain how the different eigenvalue distributions for different values of s examined in Sec. 2 influence the convergence behavior of an iterative method to solve Eq. (8).

For each of $s = -1, 0, +1$, we discretize Eq. (8) using the FDFD method for the system illustrated in Fig. 1 with an x-polarized electric dipole current source placed at the center of the simulation domain. We then solve the discretized equation by an iterative method to observe the convergence behavior. The iterative method to use in this section is the general minimal residual (GMRES) method [33], which is one of the Krylov subspace methods [5]. We use GMRES without restart because the system is sufficiently small.

Like other iterative methods, GMRES generates an approximate solution x_m of $Ax = b$ at the mth iteration step. As m increases, x_m eventually converges to the exact solution. We assume that convergence is achieved when the residual vector

$$r_m = b - Ax_m \tag{17}$$

satisfies $\|r_m\|/\|b\| < \tau$, where $\|\cdot\|$ is the 2-norm and τ is a user-defined small positive number. In practice, $\tau = 10^{-6}$ is sufficiently small for accurate solutions. We use $x_0 = 0$ as an initial guess solution throughout the paper.

Figure 3 shows $\|r_m\|/\|b\|$ versus the number m of iteration steps for the three values of s. As can be seen in the figure, the convergence behavior of GMRES is quite different for different s, with $s = -1$ far more superior than the other two choices of s.

The overall trend of the convergence behavior shown in Fig. 3 is consistent with the mathematical theories of iterative methods. For example, the convergence stagnates initially for $s = 0$, and according to [21] this is typical behavior of GMRES for a matrix with many eigenvalues close to 0 such as our A for $s = 0$ (see Fig. 2(a)). Also, the convergence is very slow for $s = +1$, and Ref. [23] argues that in general the Krylov subspace methods converge much more slowly for indefinite matrices such as our A for $s = +1$ (see Fig. 2(c)) than for definite matrices. In this section we provide a more intuitive explanation for the convergence behavior by using the residual polynomial.

We first review the residual polynomial of GMRES briefly. Suppose that $\mathscr{P}_m$ is the set of all polynomials $\tilde{p}_m$ of degree at most m such that

$$\tilde{p}_m(0) = 1. \tag{18}$$

For each $\tilde{p}_m \in \mathscr{P}_m$, we can define a column vector

$$\tilde{r}_m \equiv \tilde{p}_m(A)r_0. \tag{19}$$

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

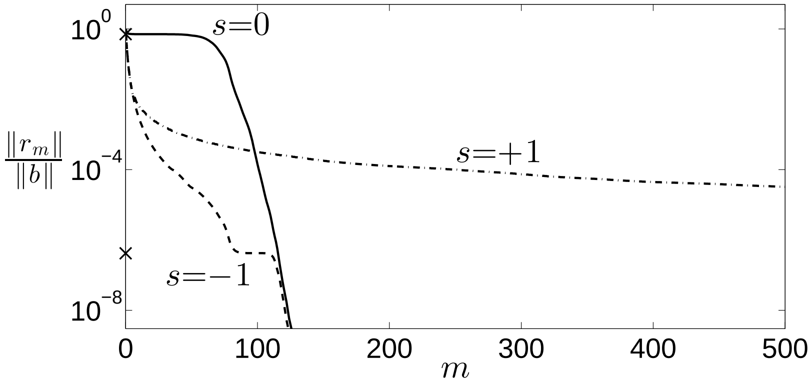


Fig. 3. Convergence behavior of GMRES for the vacuum-filled domain illustrated in Fig. 1. Three systems of linear equations discretized from Eq. (8) for $s = 0, -1, +1$ are solved by GMRES. In the iteration process of GMRES for each s, we plot the relative residual norm $\|r_m\|/\|b\|$ at each iteration step m. Notice that for $s = 0$ the relative residual norm stagnates initially; for $s = -1$ it stagnates around $m = 100$; for $s = +1$ it does not stagnate, but decreases very slowly. The upper and lower "X" marks on the vertical axis indicate the values around which our theory expects $\|r_m\|/\|b\|$ to stagnate for $s = 0$ and $s = -1$, respectively.

At the mth iteration step of GMRES, the residual vector r_m of Eq. (17) is the $\tilde{r}_m$ with the smallest 2-norm [5]. We refer to the $\tilde{p}_m$ for $\tilde{r}_m = r_m$ as the residual polynomial p_m. Therefore, from Eq. (19) we have

$$r_m = p_m(A)r_0. \tag{20}$$

Below, we show how the eigenvalue distribution of A influences p_m at each iteration step and hence influences the convergence behavior of GMRES. The matrix $A \in \mathbb{C}^{n\times n}$ for our homogeneous system described in Fig. 1 is Hermitian because it is discretized from the Hermitian operator T of Eq. (8). Hence, the eigendecomposition of A is

$$A = V\Lambda V^\dagger, \tag{21}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}, \quad V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \tag{22}$$

with real eigenvalues λ_i and the corresponding normalized eigenvectors v_i, and $V^\dagger$ is the conjugate transpose of V; note that V is unitary, i.e., $V^\dagger V = I$. Substituting Eq. (21) in Eq. (19), we obtain

$$\tilde{r}_m = V\tilde{p}_m(\Lambda)V^\dagger r_0 \tag{23}$$

because $(V\Lambda V^\dagger)^k = V\Lambda^k V^\dagger$ for any nonnegative integer k. We then define a column vector

$$\tilde{z}_m \equiv V^\dagger(\tilde{r}_m/\|b\|), \tag{24}$$

whose ith element, which is referred to as $\tilde{z}_{mi}$ below, is the projection of $\tilde{r}_m/\|b\|$ onto the direction of the ith eigenvector v_i. Similarly, we also define

$$z_m \equiv V^\dagger(r_m/\|b\|). \tag{25}$$

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

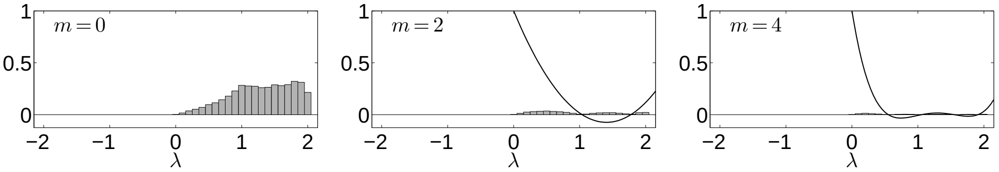


Fig. 4. Initial evolution of $r_m/\|b\|$ for $s=-1$. Relative residual vectors $r_m/\|b\|$ are visualized at three iteration steps $m=0,2,4$. In each plot, the column on each interval represents the norm of $r_m/\|b\|$ projected onto the sum of the eigenspaces of the eigenvalues contained in the interval. Notice that all the columns almost vanish only after four iteration steps. In the plots for $m=2$ and $m=4$, the residual polynomials p_m are also plotted as solid curves; note that they always satisfy the condition (18).

From Eq. (25) for $m=0$ and Eqs. (23) and (24), we obtain

$$\tilde{z}_m = \tilde{p}_m(\Lambda) z_0 = \begin{bmatrix} \tilde{p}_m(\lambda_1) & & \\ & \ddots & \\ & & \tilde{p}_m(\lambda_n) \end{bmatrix} z_0, \tag{26}$$

which can be written element-by-element as

$$\tilde{z}_{mi} = \tilde{p}_m(\lambda_i) z_{0i}. \tag{27}$$

Because $\|\tilde{z}_m\| = \|\tilde{r}_m\|/\|b\|$, GMRES minimizes $\|\tilde{z}_m\|$ to $\|z_m\|$ when it minimizes $\|\tilde{r}_m\|$ to $\|r_m\|$ at the mth iteration step.

According to Eq. (27), $|\tilde{z}_{mi}|$ is minimized to 0 when $\tilde{p}_m$ has λ_i as a root. Thus, the most ideal $\tilde{p}_m$ has all the n eigenvalues of A as its roots, because it reduces $\|\tilde{z}_m\|$ to 0. However, $\tilde{p}_m$ has at most m roots, and m, which is the number of iteration steps, is typically far less than n. Therefore, $\tilde{p}_m$ needs to have its roots optimally placed near the eigenvalues to minimize $\|\tilde{z}_m\|$. Hence, the eigenvalue distribution of A greatly influences the convergence behavior of GMRES.

We now seek to understand the convergence behavior of GMRES for the different choices of s. We begin with $s=-1$. In Fig. 4 we plot $r_m/\|b\|$ for $s=-1$ as bar graphs at the first few iteration steps. The horizontal axis in each plot represents eigenvalues. We divide the range of eigenvalues into the same 41 intervals $t_{-20},\ldots,t_0,\ldots,t_{20}$ used in Fig. 2; note that $t_0 \ni 0$. The height of the column on each interval is the norm of the projection of $r_m/\|b\|$ onto the space spanned by the eigenvectors whose corresponding eigenvalues are contained in the interval. More specifically, the height of the column on t_j after m iteration steps is

$$h_{mj} = \left[\sum\nolimits_{\lambda_i \in t_j} z_{mi}^2\right]^{1/2}. \tag{28}$$

Note that $[\sum_j h_{mj}^2]^{1/2} = \|r_m\|/\|b\|$, and thus the sum of the squares of the column heights is a direct measure of convergence.

A few properties of $r_m/\|b\|$ for $s=-1$ shown in Fig. 4 are readily predicted from the corresponding eigenvalue distribution of the matrix A presented in Fig. 2(b). For instance, A has no eigenvalues in $t_{j<0}$, and therefore $r_m/\|b\|$ has components only in $t_{j\geq 0}$ throughout the iteration process as demonstrated in Fig. 4. Also, A has very few eigenvalues in t_0, and thus $r_0/\|b\|$ has a very weak component in t_0 as can be seen in the $m=0$ plot in Fig. 4.

Now, we relate $r_m/\|b\|$ with the residual polynomial to explain the convergence behavior of GMRES for $s=-1$. The residual polynomial $p_m(\lambda)$, which is obtained by solving a least

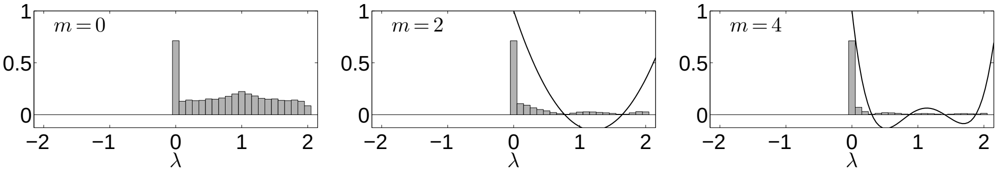


Fig. 5. Initial evolution of $r_m/\|b\|$ for $s = 0$. Relative residual vectors $r_m/\|b\|$ are visualized at three iteration steps $m = 0, 2, 4$. In each plot, the column on each interval represents the norm of $r_m/\|b\|$ projected onto the sum of the eigenspaces of the eigenvalues contained in the interval. Notice that most columns almost vanish only after four iteration steps, except for the very persistent column at $\lambda \simeq 0$. In the plots for $m = 2$ and $m = 4$, the residual polynomials p_m are also plotted as solid curves; note that they always satisfy the condition (18).

squares problem, is also plotted in Fig. 4 at each iteration step. As the iteration proceeds, the residual polynomial in Fig. 4 has more and more roots, but only in $t_{j\geq 0}$, because the eigenvalues exist only in $t_{j\geq 0}$ and the roots of residual polynomials should stay close to the eigenvalues as mentioned in the discussion following Eq. (27). Also, as Eq. (27) predicts, the columns in each plot of Fig. 4 almost vanish at the roots of the residual polynomial. Therefore, all the columns quickly shrink as the number of the roots of the residual polynomial increases in the iteration process of GMRES. The fast reduction of the column heights provides visualization of the fast convergence of GMRES for $s = -1$ shown in Fig. 3.

Next, we examine the convergence behavior for $s = 0$. Figure 5 shows $r_m/\|b\|$ for $s = 0$ at the first few iteration steps. Note that $r_0/\|b\|$ has a tall column on t_0 because A has many eigenvalues in t_0 as shown in Fig. 2(a). Also, the tall column on t_0 persists during the initial period of the iteration process.

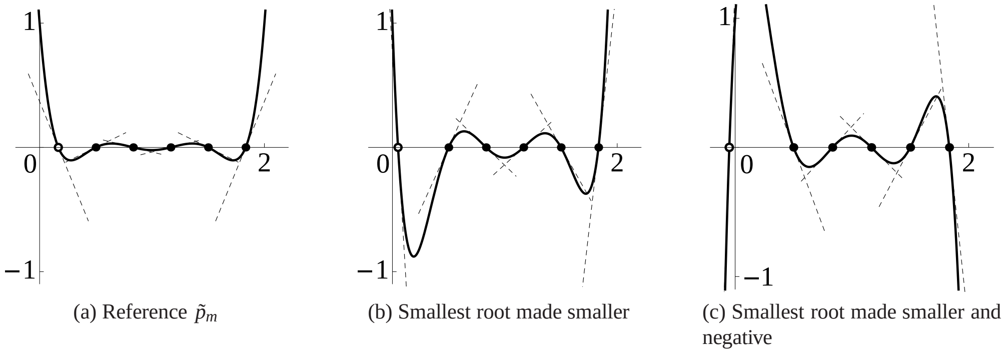
To explain the above convergence behavior for $s = 0$, we show that for a nearly positive-definite matrix the column on t_0 is persistent during the initial period of the iteration process of GMRES in general. For that purpose, we compare the three polynomials $\tilde{p}_m \in \mathscr{P}_m$ shown in Fig. 6. The three $\tilde{p}_m$ are chosen as candidates for the residual polynomial p_m for a nearly positive-definite matrix, and therefore the roots of the polynomials are placed in $t_{j\geq 0}$ according to the discussion following Eq. (27). The three $\tilde{p}_m$ have the same roots except for their smallest roots: $\tilde{p}_m$ in Fig. 6(a) does not have its smallest root in t_0, whereas $\tilde{p}_m$ in Figs. 6(b) and 6(c) do. Note that the latter two $\tilde{p}_m$ can shrink the column on t_0 more effectively than the first $\tilde{p}_m$ according to Eq. (27).

However, the slopes at the roots of the latter two $\tilde{p}_m$ are steeper than the slopes at the corresponding roots of the first $\tilde{p}_m$ as shown in Fig. 6. In Appendix B, we prove rigorously that the slopes of $\tilde{p}_m$ at all roots indeed increase as the smallest root decreases in magnitude. In general, $\tilde{p}_m$ with steeper slopes at the roots oscillates with larger amplitudes around the horizontal axis because it varies faster around the axis; compare the amplitudes of oscillation in Fig. 6(a) with those in Figs. 6(b) and 6(c). The increased amplitudes of oscillation amplify $|\tilde{z}_{mi}|$ overall according to Eq. (27), and thus $\|\tilde{z}_m\|$ as well.

In other words, shrinking the column on t_0 (by placing the smallest root of $\tilde{p}_m$ in t_0) is achieved only at the penalty of amplifying the columns on $t_{j>0}$. This penalty is too heavy when the columns on $t_{j>0}$ constitute a considerable portion of $\|\tilde{z}_m\|$. Therefore, roots of residual polynomials are not placed in t_0 until the columns on $t_{j>0}$ become quite small, which results in the persistence of the column on t_0 during the initial period of the iteration process.

Because the height of the column on t_0 remains almost the same at the initial iteration steps of GMRES, h_{00} of Eq. (28), which is the initial height of this column, provides an approximate

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013



(a) Reference $\tilde{p}_m$

(b) Smallest root made smaller

(c) Smallest root made smaller and negative

Fig. 6. Impact of the magnitude of the smallest root of a polynomial $\tilde{p}_m \in \mathscr{P}_m$ on the oscillation amplitudes of $\tilde{p}_m$. Three $\tilde{p}_m$ of degree 6 are shown. In each figure, a solid line represents a polynomial; an open dot on the horizontal axis indicates the smallest root; solid dots indicate the other roots; dashed lines show the slopes of the polynomial at the roots. The three polynomials have the same roots except for their smallest roots: the smallest root in (a) becomes smaller positive and negative roots in (b) and (c), respectively. Notice that the slopes at all roots in (a) become steeper in (b) and (c) as the smallest root decreases in magnitude, and as a result the amplitudes of oscillation of $\tilde{p}_m$ around the horizontal axis increase.

lower bound of $\|z_m\| = \|r_m\|/\|b\|$ during the initial period of the iteration process. A more accurate lower bound is calculated as the norm of $r_0/\|b\|$ projected onto the eigenspace of the eigenvalue closest to 0. For our example system, for $s = 0$ the calculated lower bound is 0.707. Note that $\|r_m\|/\|b\|$ for $s = 0$ indeed stagnates initially at this value in Fig. 3. For $s = -1$ the calculated lower bound is 4.16×10^{-7}, at which $\|r_m\|/\|b\|$ also stagnates as shown in Fig. 3. However, this value is much smaller than the lower bound for $s = 0$, because for $s = -1$ the initial height of the column on t_0 is almost negligible as shown in the $m = 0$ plot in Fig. 4. In fact, the value is smaller than the conventional tolerance $\tau = 10^{-6}$ mentioned below Eq. (17), so the stagnation does not deteriorate the convergence speed for $s = -1$.

Lastly, we examine the convergence behavior for $s = +1$. Figure 7 shows $r_m/\|b\|$ for $s = +1$ at some first ($m = 0, 4, 7, 11$) and later ($m = 120, 140$) iteration steps. Because the matrix A for $s = +1$ has both positive and negative eigenvalues as indicated in Fig. 2(c), $r_m/\|b\|$ has components in both $t_{j>0}$ and $t_{j<0}$, but in the present example the components of $r_m/\|b\|$ are concentrated in $t_{j<0}$ initially ($m = 0$ plot in Fig. 7). Thus GMRES begins with the roots of residual polynomials placed in $t_{j<0}$ ($m = 4$ plot in Fig. 7). However, such residual polynomials have large values in $t_{j>0}$, so they amplify the initially very small components of $r_m/\|b\|$ in $t_{j>0}$ according to Eq. (27), and eventually we reach a point where the components of $r_m/\|b\|$ in $t_{j>0}$ and $t_{j<0}$ become comparable ($m = 7$ plot in Fig. 7). Afterwards, GMRES places the roots of residual polynomials in both $t_{j>0}$ and $t_{j<0}$ so that the components of $r_m/\|b\|$ in both regions are reduced.

We note that the convergence behavior for $s = +1$ is initially quite similar to that for $s = -1$ because $r_0/\|b\|$ for $s = +1$ has components concentrated in $t_{j<0}$ and only a very weak component in t_0. Therefore, $\|r_m\|/\|b\|$ reduces quickly for $s = +1$ without stagnation during the initial period of the iteration process as shown in Fig. 3.

During the later period of the iteration process, however, the reduction of $\|r_m\|/\|b\|$ for $s = +1$ slows down significantly, and eventually $s = +1$ produces the slowest convergence among the three values of s as shown in Fig. 3. The slow reduction of $\|r_m\|/\|b\|$ is due to the very

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

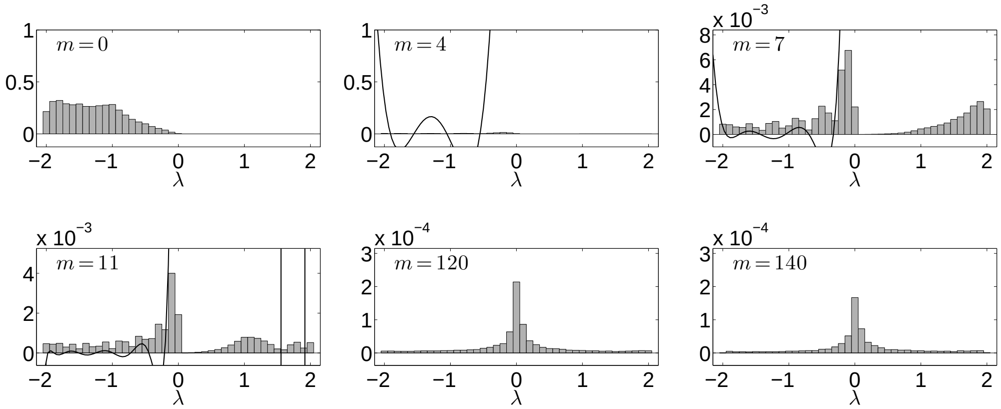


Fig. 7. Evolution of $r_m/\|b\|$ for $s=+1$. Relative residual vectors $r_m/\|b\|$ are visualized at iteration steps $m=0,4,7$ in the first row and at $m=11,120,140$ in the second row. In each plot, the column on each interval represents the norm of $r_m/\|b\|$ projected onto the sum of the eigenspaces of the eigenvalues contained in the interval. The vertical scale of the plot is magnified as the iteration proceeds. Notice that the column at $\lambda \simeq 0$ is very persistent during the later period of the iteration process ($m=120,140$). In the plots for $m=4,7,11$, the residual polynomials p_m are also plotted as solid curves; the residual polynomials are not plotted for $m=100$ and $m=120$ because they have too many roots.

persistent column on t_0: comparing the plots for $m=120$ and $m=140$ in Fig. 7 shows that the column barely reduces for 20 iteration steps.

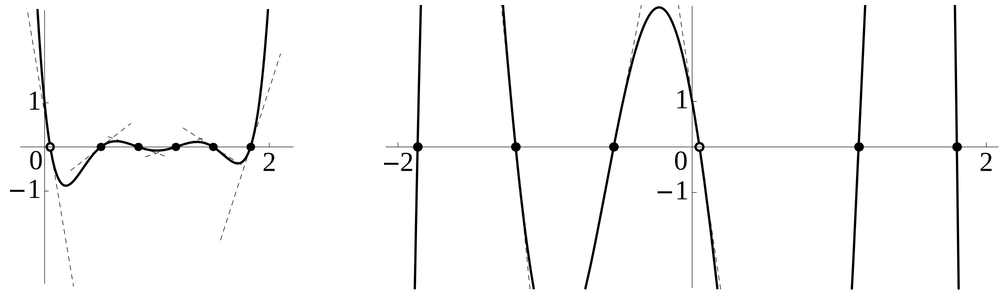
We have shown earlier that the column on t_0 is quite persistent for a nearly positive-definite matrix. The argument relied on the properties proved in Appendix B about a polynomial $\tilde{p}_m \in \mathscr{P}_m$ with only positive roots. We can easily extend the proof in the appendix to $\tilde{p}_m$ with both positive and negative roots, and then show that the column on t_0 is persistent also for a strongly indefinite matrix, which explains the slow convergence for $s=+1$ described above. However, the explanation is insufficient to explain why the convergence is *much* slower for $s=+1$ than for $s=0$ as indicated in Fig. 3.

Here, we show that the column on t_0 is in fact even more persistent for a strongly indefinite matrix than for a nearly positive-definite matrix. For that purpose, we compare the two polynomials $\tilde{p}_m \in \mathscr{P}_m$ shown in Fig. 8. As can be seen from the locations of their roots, they are candidates for the residual polynomials for different matrices: $\tilde{p}_m$ shown in Fig. 8(a) is appropriate for a nearly positive-definite matrix (referred to as A_{def} below), and $\tilde{p}_m$ shown in Fig. 8(b) is appropriate for a strongly indefinite matrix (referred to as A_{ind} below). Moreover, we choose these two $\tilde{p}_m$ to have the same smallest-magnitude root ζ_0 in t_0. Being elements of $\mathscr{P}_m$, both $\tilde{p}_m$ satisfy Eq. (18). Hence, we have $|\tilde{p}_m'(\zeta_0)| \simeq 1/|\zeta_0|$ for both $\tilde{p}_m$, where $\tilde{p}_m'$ is the first derivative of $\tilde{p}_m$.

Now, we note that $|\tilde{p}_m'|$ evaluated at a root of $\tilde{p}_m$ tends to decrease as the root gets closer to the median of the roots; see Appendix C for a more rigorous explanation. Hence, $|\tilde{p}_m'| \leq 1/|\zeta_0|$ tends to hold at most roots of $\tilde{p}_m$ for A_{def}, because ζ_0 is one of the farthest roots from the median of the roots. On the other hand, $|\tilde{p}_m'| \geq 1/|\zeta_0|$ tends to hold at most roots of $\tilde{p}_m$ for A_{ind}, because ζ_0 is one of the closest roots to the median of the roots. Therefore, $\tilde{p}_m$ for A_{ind} has much steeper slopes at most roots than $\tilde{p}_m$ for A_{def} in general, and thus has larger amplitudes of oscillation around the horizontal axis, as demonstrated in Fig. 8.

Combined with Eq. (27), the above argument shows that shrinking the column on t_0 (by placing the smallest-magnitude root of $\tilde{p}_m$ in t_0) increases $\|z_m\|$ much more for a strongly in-

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013



(a) $\tilde{p}_m$ for a nearly positive-definite matrix

(b) $\tilde{p}_m$ for a strongly indefinite matrix

Fig. 8. Candidates for the residual polynomials for (a) a nearly positive-definite matrix and (b) strongly indefinite matrix. In each figure, a solid line represents a polynomial $\tilde{p}_m \in \mathscr{P}_m$; an open dot on the horizontal axis indicates the smallest-magnitude root; solid dots indicate the other roots; dashed lines show the slopes of the polynomial at the roots. The two polynomials have the same smallest-magnitude root ζ_0, and thus have approximately the same slope $-1/\zeta_0$ at their smallest-magnitude roots. Note that for both $\tilde{p}_m$ the slopes get steeper at the roots further away from the median of the roots. Hence, the slopes of most dashed lines are gentler than $1/|\zeta_0|$ in (a) and steeper than $1/|\zeta_0|$ in (b). As a result, $\tilde{p}_m$ in (b) has larger amplitudes of oscillation around the horizontal axis than $\tilde{p}_m$ in (a).

definite matrix than for a nearly positive-definite matrix. Therefore, the column on t_0 should be much more persistent for a strongly indefinite matrix than for a nearly positive-definite matrix in general, which explains the much slower convergence for $s = +1$ than for $s = 0$ in Fig. 3.

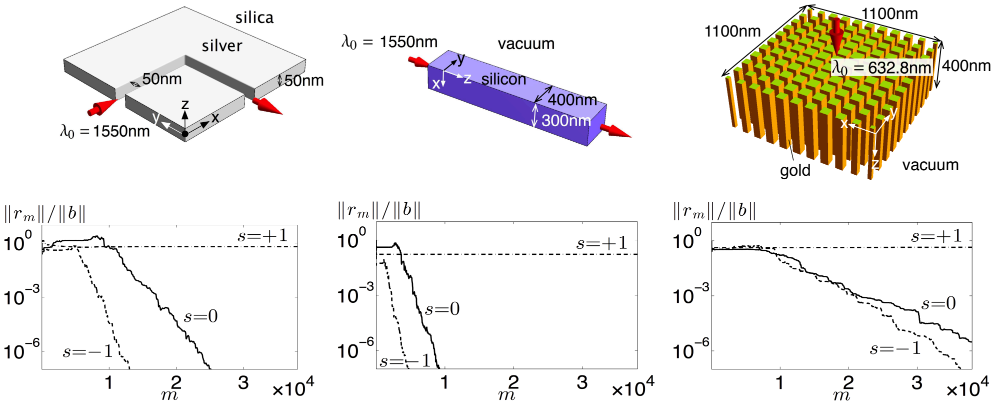
In summary of this section, we have shown that $s = -1$ produces the most superior convergence behavior; $s = 0$ induces stagnation during the initial period of the iteration process due to the high multiplicity of eigenvalues near zero; $s = +1$ leads to the slowest convergence overall due to the strongly indefinite matrix. We have provided a graphical explanation of the difference in the convergence behavior of GMRES, for which a strong theoretical basis exists, using a simple system of a homogeneous dielectric medium.

The arguments provided in this section can be easily extended to show that $s = -1$ is indeed optimal among all values in general. Compared with the case of $s = -1$, according to Eq. (14), for $s > 0$ A is always more strongly indefinite and therefore the convergence should be slower; for $-1 < s < 0$ A has more eigenvalues clustered near zero and thus the initial stagnation period should be longer; for $s < -1$ A has a wider eigenvalue value range, so the condition number of A should be larger and the convergence should be slower [23]. In other words, $s = -1$ is the value that leaves A nearly positive-definite while removing the eigenvalues clustered near zero sufficiently without increasing the condition number. With separate numerical experiments we have verified that $s = -1$ is indeed superior to values other than $s = 0$ and $s = +1$ as well.

In the next section we will see that the difference in the convergence behavior for different choices of s is in fact quite general in practical situations.

4. Convergence behavior of QMR for 3D inhomogeneous systems

In this section, we solve Eq. (7) for 3D inhomogeneous systems of practical interest by an iterative method, and demonstrate that $s = -1$ still induces faster convergence than $s = 0$ and $s = +1$. We note that the systems examined in this section are inhomogeneous and have complex ε in general. The analyses in Secs. 2 and 3, therefore, do not hold strictly here. Nevertheless, we will see that the analyses for the homogeneous system in the previous sections provide insight



(a) Slot (b) Diel (c) Array

Fig. 9. Three inhomogeneous systems for which Eq. (7) is solved for $s = -1, 0, +1$ by QMR: (a) a slot waveguide bend formed in a thin silver film (Slot), (b) a straight silicon waveguide (Diel), and (c) an array of gold pillars (Array). The figures in the first row describe the three systems. The directions of wave propagation are shown by red arrows, beside which the vacuum wavelengths used are indicated. For all three systems, the waves are excited by electric current sources **J** strictly inside the simulation domain. The plots in the second row show the convergence behavior of QMR. Note that for all three systems QMR converges fastest for $s = -1$, whereas it barely converges for $s = +1$. The relative electric permittivities of the materials used in these systems are $\varepsilon_r^{\text{silver}} = -129 - i3.28$ [34], $\varepsilon_r^{\text{silica}} = 2.085$ [35], $\varepsilon_r^{\text{silicon}} = 12.09$ [35], and $\varepsilon_r^{\text{gold}} = -10.78 - i0.79$ [36], respectively.

Table 2. Specification of the finite-difference grids used for the three systems in Fig. 9. Slot uses a nonuniform grid with smoothly varying grid cell size. The matrix A has $3N_xN_yN_z$ rows and columns, where the extra factor 3 accounts for the three Cartesian components of the E-field.

	Slot	Diel	Array
N_x, N_y, N_z	$192, 192, 240$	$220, 220, 320$	$220, 220, 130$
$\Delta_x, \Delta_y, \Delta_z$	$2 \sim 20\,\text{nm}$	$10\,\text{nm}$	$5, 5, 20\,\text{nm}$

in understanding the convergence behavior for more general systems examined in this section.

The three 3D inhomogeneous systems we consider are illustrated in the first row of Fig. 9. To prevent reflection of EM waves from boundaries, we enclose each system by the stretched-coordinate perfectly matched layer (SC-PML), because SC-PML produces a much better-conditioned matrix than the more commonly used uniaxial PML (UPML) [32]. For each system, we construct three systems of linear equations $Ax = b$ corresponding to $s = -1, 0, +1$ by the FDFD method. The number of the grid cells N_x, N_y, and N_z and the grid cell sizes Δ_x, Δ_y, and Δ_z of the finite-difference grid used to discretize each system are shown in Table 2. Considering the parameters summarized in Table 3 and the condition (16), all the three systems are in the low-frequency regime.

The constructed systems of linear equations are solved by the quasi-minimal residual (QMR) method [37], which is another Krylov subspace method. GMRES that was used in Sec. 3 to solve a 2D problem is not suitable for 3D problems here because it requires too much memory [5].

The second row of Fig. 9 shows the convergence behavior of QMR for the three systems.

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

Table 3. Parameters used in Eq. (16) for the three systems in Fig. 9. When substituted in Eq. (16), these parameters prove that all the three systems are in the low-frequency regime.

	Slot	Diel	Array
λ_0	$1550\,\text{nm}$	$1550\,\text{nm}$	$632.8\,\text{nm}$
Δ_{min}	$2\,\text{nm}$	$10\,\text{nm}$	$5\,\text{nm}$
$\max\lvert\varepsilon_r\rvert$	129.0	12.09	10.81

Note that for all three systems the choice of $s = -1$ results in the fastest convergence, and the choice of $s = +1$ barely leads to convergence. The three systems shown in Fig. 9 are chosen deliberately to include different materials such as dielectrics and metals and geometries with different degrees of complexity. Therefore, Fig. 9 suggests that the superiority of $s = -1$ over both $s = 0$ and $s = +1$ is quite general.

Even though both the iterative method and the systems in this section are significantly different from those in the previous section, the convergence behaviors are very similar. We explain the resemblance as follows.

The matrix for an inhomogeneous system is actually not much different from that for a homogenous system in many cases. Indeed, most inhomogeneous systems consist of several homogeneous subdomains. Inside each homogeneous subdomain of such an inhomogeneous system, the differential operator in Eq. (7) for the inhomogeneous system is the same as the differential operator (9) for a homogeneous system, whereas at the interfaces between the subdomains it is not. Nevertheless, the number of finite-difference grid points assigned at the interfaces is usually much smaller than that of the grid points assigned inside the homogeneous subdomains. Therefore most rows of the matrix for the inhomogeneous system should be the same as those for a homogeneous system discretized on the same grid.

In addition, the differential operator (9) for a homogeneous system is nearly Hermitian in the low-frequency regime even if ε is complex, because it is approximated well by the Hermitian operator (10).

Hence, the matrix for an inhomogeneous system is actually quite similar to the nearly Hermitian matrix for a homogeneous system. Because QMR reduces to GMRES for Hermitian matrices in exact arithmetic [38], it is reasonable that the convergence behavior of QMR for an inhomogeneous system is similar to that of GMRES for a homogeneous system.

The matrix for an inhomogeneous system deviates more from that for a homogeneous system as the number of homogeneous subdomains increases, because then the number of grid points assigned at the interfaces between homogeneous subdomains increases. Therefore, we can expect that the convergence behavior for an inhomogeneous system would deviate from that for a homogeneous system as the number of homogeneous subdomains increases. Such deviation is demonstrated in Fig. 9(c), where the system has many metallic pillars; note that the convergence behavior for $s = -1$ is no longer very different from that for $s = 0$ in this case.

5. Conclusion and final remarks

We have introduced a new method to accelerate the convergence of iterative solvers of the frequency-domain Maxwell's equations in the low-frequency regime. The method solves a new equation that is modified from the original Maxwell's equations using the continuity equation.

The operator of the newly formulated equation does not have the high multiplicity of near-zero eigenvalues that makes the convergence stagnate for the original operator. At the same time, the new operator is nearly positive-definite, so it avoids the long-term slow convergence that indefinite operators suffer from.

In this paper, we have considered only nonmagnetic materials ($\mu = \mu_0$). For magnetic mate-

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

rials ($\mu \neq \mu_0$), we note that a similar equation

$$\nabla \times \mu^{-1} \nabla \times \mathbf{E} + s\nabla \left[(\mu\varepsilon)^{-1} \nabla \cdot (\varepsilon \mathbf{E}) \right] - \omega^2 \varepsilon \mathbf{E} = -i\omega \mathbf{J} + s\frac{i}{\omega} \nabla \left[(\mu\varepsilon)^{-1} \nabla \cdot \mathbf{J} \right], \tag{29}$$

which can also be formulated from Maxwell's equations and the continuity equation, can be used instead of Eq. (7) to accelerate the convergence of iterative methods. We leave the discussion on the optimal value of s in this equation for future work.

Because our method achieves accelerated convergence by formulating a new equation, it can be easily combined with other acceleration techniques such as preconditioning and better iterative methods.

Appendix A: Eigenvalues and eigenfunctions of $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$

Using the **k**-space representations of the operators, in this section we derive the eigenvalues Eq. (12) of $\nabla \times (\nabla \times \quad)$ and Eq. (13) of $\nabla(\nabla \cdot \quad)$ as well as their corresponding eigenfunctions.

Because both $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$ are translationally invariant, their eigenfunctions have the form [39]

$$\mathbf{F} = \mathbf{F}_{\mathbf{k}} e^{-i\mathbf{k}\cdot\mathbf{r}}, \tag{30}$$

where $\mathbf{r}$ represents position, $\mathbf{k} = \hat{\mathbf{x}}k_x + \hat{\mathbf{y}}k_y + \hat{\mathbf{z}}k_z$ is a real constant wavevector, and $\mathbf{F}_{\mathbf{k}} = \hat{\mathbf{x}}F_{\mathbf{k}}^x + \hat{\mathbf{y}}F_{\mathbf{k}}^y + \hat{\mathbf{z}}F_{\mathbf{k}}^z$ is a **k**-dependent vector.

We reformulate the eigenvalue equations $\nabla \times (\nabla \times \mathbf{F}) = \lambda \mathbf{F}$ and $\nabla(\nabla \cdot \mathbf{F}) = \lambda \mathbf{F}$ by substituting Eq. (30) for $\mathbf{F}$. Then, the eigenvalue equation for $\nabla \times (\nabla \times \quad)$ is

$$\begin{bmatrix} k_y^2 + k_z^2 & -k_x k_y & -k_x k_z \\ -k_y k_x & k_z^2 + k_x^2 & -k_y k_z \\ -k_z k_x & -k_z k_y & k_x^2 + k_y^2 \end{bmatrix} \begin{bmatrix} F_{\mathbf{k}}^x \\ F_{\mathbf{k}}^y \\ F_{\mathbf{k}}^z \end{bmatrix} = \lambda \begin{bmatrix} F_{\mathbf{k}}^x \\ F_{\mathbf{k}}^y \\ F_{\mathbf{k}}^z \end{bmatrix}, \tag{31}$$

and the eigenvalue equation for $\nabla(\nabla \cdot \quad)$ is

$$-\begin{bmatrix} k_x^2 & k_x k_y & k_x k_z \\ k_y k_x & k_y^2 & k_y k_z \\ k_z k_x & k_z k_y & k_z^2 \end{bmatrix} \begin{bmatrix} F_{\mathbf{k}}^x \\ F_{\mathbf{k}}^y \\ F_{\mathbf{k}}^z \end{bmatrix} = \lambda \begin{bmatrix} F_{\mathbf{k}}^x \\ F_{\mathbf{k}}^y \\ F_{\mathbf{k}}^z \end{bmatrix}. \tag{32}$$

By solving Eqs. (31) and (32) for a given $\mathbf{k}$, we obtain

$$\lambda = 0,\ |\mathbf{k}|^2,\ |\mathbf{k}|^2, \tag{33}$$

which is Eq. (12), as the eigenvalues of $\nabla \times (\nabla \times \quad)$, and

$$\lambda = -|\mathbf{k}|^2,\ 0,\ 0, \tag{34}$$

which is Eq. (13), as the eigenvalues of $\nabla(\nabla \cdot \quad)$, and Eq. (30) with

$$\mathbf{F}_{\mathbf{k}} = \begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix}, \begin{bmatrix} k_z \\ 0 \\ -k_x \end{bmatrix}, \begin{bmatrix} -k_y \\ k_x \\ 0 \end{bmatrix} \tag{35}$$

as the eigenfunctions corresponding to both Eqs. (33) and (34).

We note from Eqs. (33) and (34) that $\nabla \times (\nabla \times \quad)$ and $\nabla(\nabla \cdot \quad)$ are positive-semidefinite and negative-semidefinite, respectively.

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

Appendix B: Effect of the smallest root of $\tilde{p}_m \in \mathscr{P}_m$ on the slopes at the roots

In this section, we show that the slopes at the roots of a polynomial $\tilde{p}_m \in \mathscr{P}_m$ with all positive roots become steeper when the smallest root decreases in magnitude. This behavior is illustrated in Fig. 6.

Since $\tilde{p}_m \in \mathscr{P}_m$ satisfies the condition (18), it can be factored as

$$\tilde{p}_m(\zeta) = \prod_{i=1}^{d_m} \left(1 - \frac{\zeta}{\zeta_i}\right), \tag{36}$$

where $d_m \le m$ is the degree of $\tilde{p}_m$ and ζ_i's are the roots of $\tilde{p}_m$. Hence, the slope of $\tilde{p}_m$ at a root ζ_k is

$$\tilde{p}_m'(\zeta_k) = -\frac{1}{\zeta_k} \prod_{i \neq k} \left(1 - \frac{\zeta_k}{\zeta_i}\right). \tag{37}$$

Now, suppose that $0 < \zeta_1 < \cdots < \zeta_{d_m}$. We can easily show that $|\tilde{p}_m'(\zeta_k)|$ increases for any k when ζ_1 decreases toward zero (while remaining positive) as follows. For $k = 1$, we have

$$\left|\tilde{p}_m'(\zeta_1)\right| = \frac{1}{\zeta_1} \left(1 - \frac{\zeta_1}{\zeta_2}\right) \cdots \left(1 - \frac{\zeta_1}{\zeta_{d_m}}\right), \tag{38}$$

which clearly increases as ζ_1 decreases to 0. For $k > 1$, we have

$$\left|\tilde{p}_m'(\zeta_k)\right| = \left(\frac{\zeta_k}{\zeta_1} - 1\right) \left[\frac{1}{\zeta_k} \prod_{i \neq 1,k} \left|1 - \frac{\zeta_k}{\zeta_i}\right|\right], \tag{39}$$

where the parentheses enclose the only quantity that is dependent on ζ_1. We can therefore see that $|\tilde{p}_m'(\zeta_k)|$ increases as ζ_1 decreases for $k > 1$ as well. Therefore, for a given $\tilde{p}_m \in \mathscr{P}_m$ whose roots are all positive, the slopes of $\tilde{p}_m$ at the roots become steeper if the smallest root decreases in magnitude while remaining positive. This situation is illustrated by the transition from Fig. 6(a) to Fig. 6(b).

The slopes at the roots also become steeper when the originally positive ζ_1 is replaced by a negative value, as long as the negative value is smaller in magnitude than the original ζ_1. Replacing the originally positive ζ_1 with a negative quantity that is smaller in magnitude is equivalent to first replacing ζ_1 with a smaller positive value and then flipping its sign. Because we have already shown above that the slopes get steeper when the originally positive ζ_1 is replaced by a smaller positive value, it is sufficient to show that flipping the sign of ζ_1 makes the slopes even steeper. For a negative ζ_1, the slopes at the roots are

$$\left|\tilde{p}_m'(\zeta_1)\right| = \frac{1}{|\zeta_1|} \left(1 + \frac{|\zeta_1|}{\zeta_2}\right) \cdots \left(1 + \frac{|\zeta_1|}{\zeta_{d_m}}\right) \tag{40}$$

and

$$\left|\tilde{p}_m'(\zeta_k)\right| = \left(\frac{\zeta_k}{|\zeta_1|} + 1\right) \left[\frac{1}{\zeta_k} \prod_{i \neq 1,k} \left|1 - \frac{\zeta_k}{\zeta_i}\right|\right] \tag{41}$$

for $k > 1$. These slopes are steeper than Eqs. (38) and (39), respectively, which are the slopes for a positive ζ_1 with the same magnitude. Therefore, for a given $\tilde{p}_m \in \mathscr{P}_m$ whose roots are all positive, the slopes of $\tilde{p}_m$ at the roots become steeper if the smallest root is replaced by the one that is smaller in magnitude but negative. This situation is illustrated by the transition from Fig. 6(a) to Fig. 6(c).

#194688 - $15.00 USD Received 26 Jul 2013; revised 8 Sep 2013; accepted 9 Sep 2013; published 18 Sep 2013

Appendix C: Trend in the slopes of a polynomial at the roots

In this section, we consider a polynomial p with all real roots, and show that the slope of p evaluated at a root closer to the median of the roots tends to be gentler than the slope evaluated at a root farther away from the median of the roots. This behavior is illustrated in Fig. 8.

Consider a polynomial of degree m,

$$p(\zeta) = \alpha \prod_{i=1}^{m} (\zeta - \zeta_i), \tag{42}$$

with $\zeta_1 < \cdots < \zeta_m$. The slope of p at a root ζ_k is

$$p'(\zeta_k) = \alpha \prod_{i \neq k} (\zeta_k - \zeta_i). \tag{43}$$

Now, we evaluate $|p'(\zeta_{k+1})|/|p'(\zeta_k)|$. We first consider the case where the roots are evenly spaced, i.e., $\zeta_{i+1} - \zeta_i = \text{(const.)}$, for which we have

$$\frac{|p'(\zeta_{k+1})|}{|p'(\zeta_k)|} = \frac{k!\,(m-k-1)!}{(k-1)!\,(m-k)!} = \frac{k}{m-k}. \tag{44}$$

Equation (44) is an increasing function of k for $1 \leq k \leq m-1$, and it is less than 1 for $k < m/2$ and greater than 1 for $k > m/2$. Therefore, $|p'(\zeta_k)|$ is largest for $k = 1$ and $k = m$, and it decreases as k becomes closer to $k = \lfloor (m+1)/2 \rfloor$ and $k = \lceil (m+1)/2 \rceil$, which are the medians of the indices. In other words, for p with evenly spaced roots, the slopes of p get gentler at the roots closer to the median of the roots.

It is reasonable to expect that the above trend in the slopes also holds for p with unevenly spaced roots, unless the unevenness is too severe. To verify the expectation, we examine $|p'(\zeta_{k+1})|/|p'(\zeta_k)|$ without assuming $\zeta_{i+1} - \zeta_i = \text{(const.)}$:

$$\begin{aligned} \frac{|p'(\zeta_{k+1})|}{|p'(\zeta_k)|} &= \frac{\prod_{i \neq k+1} |\zeta_{k+1} - \zeta_i|}{\prod_{i \neq k} |\zeta_k - \zeta_i|} = \prod_{i \neq k, k+1} \frac{|\zeta_{k+1} - \zeta_i|}{|\zeta_k - \zeta_i|} = \prod_{i=1}^{k-1} \left(\frac{\zeta_{k+1} - \zeta_i}{\zeta_k - \zeta_i} \right) \prod_{i=k+2}^{m} \left(\frac{\zeta_i - \zeta_{k+1}}{\zeta_i - \zeta_k} \right) \\ &= \left[\prod_{i=1}^{k-1} \left(1 + \frac{\zeta_{k+1} - \zeta_k}{\zeta_k - \zeta_i} \right) \right] \left[\prod_{i=k+2}^{m} \left(1 - \frac{\zeta_{k+1} - \zeta_k}{\zeta_i - \zeta_k} \right) \right]. \end{aligned} \tag{45}$$

Here, the factors within the first (second) brackets are always greater (less) than 1, so the number of factors greater (less) than 1 increases (decreases) for increasing k. Hence, $|p'(\zeta_{k+1})|/|p'(\zeta_k)|$ tends to be less than 1 for smaller k, and it tends to be greater than 1 for larger k. This means that as k increases $|p'(\zeta_k)|$ tends to decrease first and then tends to increase. Therefore, even if the roots of p are unevenly spaced, the slopes of p tend to get gentler at the roots closer to the median of the roots.

Acknowledgment

We thank Prof. Michael Saunders for helpful discussions about iterative methods. This work was supported by the National Science Foundation (Grant No. DMS-0968809), the AFOSR MURI program (Grant No. FA9550-09-1-0704), and the Interconnect Focus Center, funded under the Focus Center Research Program, a Semiconductor Research Corporation entity. Wonseok Shin also acknowledges the support of Samsung Scholarship.