

# THEORETICAL DESCRIPTION OF THE EMFLEX FINITE ELEMENT FORMULATION

September 1991

## §1.0 Introduction

The purpose of this theoretical section is to derive and analyze three-dimensional finite element formulations of Maxwell's equations governing classical electromagnetic propagation in dielectrics. The derivation provides a comprehensive analytical description of the finite element equations, while the analysis quantifies accuracy of time-harmonic plane wave propagation as a function of wave direction and discretization. The derivation is limited to so-called cartesian elements (hexahedrons) for reasons of modeling simplicity and computational efficiency. Skewed elements, although an important attribute of the finite element method, are not discussed for the sake of brevity.

A tacit assumption in our formulation is that electromagnetic waves are supported throughout an infinite 3D space. In other words, there is no outer boundary limiting the field. Clearly, however, solving infinite-domain field problems is not practical. Therefore, 3D space must be divided into a finite interior or computational domain and an infinite exterior domain. By hypothesis, the interior includes all physical features of interest while the exterior is effectively "featureless" so that it propagates energy outward with negligible backscatter. This artifice allows replacement of the exterior domain with a so-called radiation boundary condition on the outer surface of the interior domain, and defines the boundary value part of our problem.

The finite element formulation of initial-boundary value problems governed by time-dependent partial differential equations (PDEs)—Maxwell's equations in particular—consists of the following formal steps:

- 1) Partition the problem's interior domain into a number of logically regular, contiguous subdomains, i.e., the "model";
- 2) Represent the field over each subdomain by a simplified basis function that interpolates between discrete field points or nodes;
- 3) Convert the point-wise partial differential operator to an equivalent but "weaker" scalar integral operator admitting lower order derivatives;
- 4) Evaluate the integral operator for the simplified field basis, giving an algebraic system of equations on the nodal field vector and its time derivatives;
- 5) Apply a radiation condition on the interior domain's boundary in order to simulate scattering into the infinite exterior domain.
- 6) Solve the system of ordinary differential equations (ODEs) in time using finite differences, modal analysis, etc.

The finite element part, steps 1 to 4, yields an approximate integration of the PDE's spatial differential operator. In other words, the so-called finite element discretization is nothing more than a quadrature formula. The remaining 1D temporal problem, step 6, is typically integrated in a more conventional fashion.

Formal reduction of the point-wise partial differential equation to a finite element form may be accomplished in at least two ways—the method of weighted residuals (Galerkin’s method) and a variational principle. They are fundamentally equivalent for self-adjoint (symmetric) differential operators provided that consistent assumptions are made, although the method of weighted residuals is more general. The approach chosen usually depends on the analyst’s perspective or background, e.g., mathematical, engineering, etc., rather than any compelling analytical reason. A somewhat simplified derivation will be given here that is comprehensive yet minimizes nomenclature and historical biases. For further reading, a succinct description of the formalism in the context of linear differential operators may be found in Zienkiewicz (1985).

The most difficult part of the finite element formulation of propagation-type problems is deriving an effective radiation or absorbing boundary condition. This is an approximate condition on the exterior boundary of the finite element model that discriminates between incident (illumination) and scattered radiation and selectively absorbs the scattered part, mimicking radiation into an infinite, nonreflecting exterior domain. An effective condition that is sufficient for simultaneous plane wave illumination and scattering is described below

The derivation begins with a description of Maxwell’s equations in §2.0 sufficient for our purposes. In §3.0 the weak form of Maxwell’s equations is derived. In §4.0 an assumption is made on the mathematical form of solutions to produce the global ordinary differential equations, i.e., the familiar finite element system of equations. In §5.0 the system of equations is specialized to particular shape functions and finite element types to yield the working “elemental” equations. In §6.0 these exact elemental equations are assembled and their accuracy is quantified in the frequency domain as a function of wavelength. In §7.0 an approximate elemental equation that exhibits superior performance is described and accuracy quantified; this is the basis for the actual implementation in EMFlex. In §8.0 the problem of radiation boundary conditions is considered, including graphical comparison of various approaches. Finally, in §9.0 the issues of time-domain versus frequency-domain solution methods are discussed and quantitative comparisons of these approaches are presented. §10 is a list of references.

## §2.0 Maxwell’s Equations

Maxwell’s equations provide the mathematical basis for rigorous analysis of classical electromagnetic wave propagation. In particular, they provide a complete description of macroscopic optical phenomena in dielectric media. There are lower limits on size and intensity where quantum behavior becomes significant, but for nearly all scales of practical interest, Maxwell’s equations are both necessary and sufficient. This section presents various forms of the equations along with ancillary relations that are useful for numerical algorithm development.

Maxwell originally proposed an arcane system of 20 equations in 20 unknowns. The system was subsequently simplified by Heaviside and Hertz to its modern form, namely

$$-\dot{\mathbf{B}} = \nabla \times \mathbf{E} \quad , \quad \dot{\mathbf{D}} + \mathbf{J} = \nabla \times \mathbf{H} \quad (2.1)$$

where  $\mathbf{B}$  is magnetic induction,  $\mathbf{E}$  is electric field intensity,  $\mathbf{D}$  is electric displacement,  $\mathbf{J}$  is current density, and  $\mathbf{H}$  is magnetic field intensity. Continuity of  $\mathbf{E}$  and  $\mathbf{H}$  is required in order to define the spatial derivatives. Note that in (2.1) bold letters represent vectors,  $\nabla \times$  is the curl operator, and time derivatives are denoted by a dot above the variable.

A useful alternative to Maxwell's point-wise partial differential equations are volumetric forms obtained by integrating (2.1) over space. Applying the so-called curl theorem—analogous to the divergence theorem of Gauss—to the resulting integrals of  $\nabla \times \mathbf{E}$  and  $\nabla \times \mathbf{H}$  yields the vector integral equations

$$-\int_{\Omega} \dot{\mathbf{B}} d\Omega = \int_{\Sigma} \mathbf{n} \times \mathbf{E} d\Sigma \quad , \quad \int_{\Omega} (\dot{\mathbf{D}} + \mathbf{J}) d\Omega = \int_{\Sigma} \mathbf{n} \times \mathbf{H} d\Sigma \quad (2.2)$$

where  $\Omega$  is the domain of integration,  $\Sigma$  is its boundary, and  $\mathbf{n}$  is the outward unit normal vector to  $\Sigma$ . This form removes the restriction on field continuity since spatial derivatives no longer appear. More conventional scalar integral equations may be written by applying Stokes' theorem to Maxwell's equations, yielding the famous laws of Ampere and Faraday.

To make Maxwell's equations determinate for  $\mathbf{B}$ ,  $\mathbf{E}$ ,  $\mathbf{D}$ ,  $\mathbf{J}$ , and  $\mathbf{H}$ , so-called constitutive relations must be defined. In nearly all cases the linear relations,

$$\mathbf{B} = \mu \mathbf{H} \quad , \quad \mathbf{D} = \epsilon \mathbf{E} \quad , \quad \mathbf{J} = \sigma \mathbf{E} \quad (2.3)$$

suffice. Proportionality factors,  $\mu$ ,  $\epsilon$ , and  $\sigma$ , are magnetic permeability, dielectric permittivity, and conductivity, respectively. Provided the medium is isotropic, these factors are scalars, otherwise they are tensors. Substituting (2.3) into (2.1) gives the determinate form of Maxwell's partial differential equations

$$-\mu \dot{\mathbf{H}} = \nabla \times \mathbf{E} \quad , \quad \epsilon \dot{\mathbf{E}} + \sigma \mathbf{E} = \nabla \times \mathbf{H} \quad (2.4)$$

relating the time rate of change of the magnetic field to the curl (vorticity or "swirl") of the electric field and vice versa. These assume that  $\epsilon$  and  $\mu$  vary with time slowly, if at all, in comparison to the fields themselves. For nonmagnetic materials permeability  $\mu$  is essentially equal to its vacuum value,  $\mu_0$ , everywhere.

Since  $\mathbf{E}$  is the primary field unknown for analyses in dielectric media,  $\mathbf{H}$  may be eliminated between the two curl equations in (2.4) and treated as a secondary or derived quantity. Taking the time derivative of the second equation in (2.4), the curl of the first, and eliminating the term with  $\mathbf{H}$  gives the second order partial differential equation,

$$\epsilon \ddot{\mathbf{E}} + \sigma \dot{\mathbf{E}} = -\frac{1}{\mu} \nabla \times \nabla \times \mathbf{E} \quad (2.5)$$

where the constant magnetic permeability is brought outside the curl operator. Such a simple equation cannot be written for  $\mathbf{H}$  because  $\epsilon$ , unlike  $\mu$ , is a function of position and its gradient must be included. Using the vector identity,  $\nabla \times \nabla \times \mathbf{E} = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}$ , (2.5) may also be written

$$\epsilon \ddot{\mathbf{E}} + \sigma \dot{\mathbf{E}} = \frac{1}{\mu} \nabla^2 \mathbf{E} - \frac{1}{\mu} \nabla(\nabla \cdot \mathbf{E}) \quad (2.6)$$

Note the similarity between this equation and Navier's equation

$$\rho \ddot{\mathbf{U}} = G \nabla^2 \mathbf{U} - (\lambda + G) \nabla (\nabla \cdot \mathbf{U}) \quad (2.7)$$

describing displacement  $\mathbf{U}$  in a linear elastic medium. As physicists finally accepted in the late 1800s, this similarity is in appearance only. Nonetheless, it helps in solving the electromagnetic equations since many of the numerical algorithms developed for elasticity over the last 30 years are directly applicable to (2.5,6).

It should be noted for completeness that the vector fields in (2.1) are ultimately caused by some distribution of electric charge,  $\rho$ , and current,  $\mathbf{J}$ , generally related by the continuity equation,

$$\nabla \cdot \mathbf{J} + \dot{\rho} = 0 \quad (2.7)$$

expressing point-wise conservation of charge. Taking the divergence of (2.1), substituting (2.7), and integrating over time, assuming a quiescent initial or final state, gives the divergence conditions,

$$\nabla \cdot \mathbf{B} = 0 \quad , \quad \nabla \cdot \mathbf{D} = \rho \quad (2.8)$$

These are often appended to Maxwell's equations but are dependent conditions, either in part ( $\nabla \cdot \mathbf{B} = 0$ ), or wholly if charge is conserved ( $\nabla \cdot \mathbf{D} = \rho$ ). Observe that by solving  $\nabla \cdot \mathbf{D} = \nabla \cdot \epsilon \mathbf{E} = \rho$  for  $\nabla \cdot \mathbf{E} = \rho / \epsilon - \nabla \epsilon / \epsilon \cdot \mathbf{E}$  and replacing  $\nabla \cdot \mathbf{E}$  in (2.6), a second order equation results that implicitly incorporates charge conservation.

### §3.0 A Weak Form of the Electric Field Equation

To apply the conventional finite element formalism to Maxwell's equations, it is convenient to start with the second order PDE on electric field, (2.5), rather than the original system of first order equations, (2.4). Strict solutions of this equation must possess at least second derivatives, however, it is impractical to require such continuity from numerical approximations. A better approach is to rewrite the equation in an integral form admitting lower order derivatives. This is the so-called weak formulation.

To derive the weak form of (2.5) it is necessary to define another field over the wave domain, the so-called test function,  $\mathbf{G}(\mathbf{x}, t)$ . This is a completely arbitrary function within wave domain  $\Omega$ . Taking the inner (dot) product of (2.5) with  $\mathbf{G}$  and integrating over  $\Omega$  gives

$$\int_{\Omega} \mathbf{G} \cdot (\epsilon \ddot{\mathbf{E}} + \sigma \mathbf{E}) d\Omega \equiv - \int_{\Omega} \mathbf{G} \cdot \frac{1}{\mu} \nabla \times \nabla \times \mathbf{E} d\Omega \quad (3.1)$$

Multiplication by a test function and integration reduces the point-wise vector equation to a volumetric scalar equation—the weak form. It is easy to prove the assertion that if this integral equation is satisfied for any  $\mathbf{G}$  then the PDE is necessarily satisfied at all points in the domain. The converse is certainly true, but if the PDE is not satisfied in some subdomain then a test function can be chosen that makes the integral nonzero, hence the assertion is true.

Consider the right-hand integrand in (3.1), i.e.,  $\mathbf{G} \cdot \nabla \times (\nabla \times \mathbf{E})$  after factoring  $\mu$ . From the vector identity,  $\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}) - \mathbf{A} \cdot (\nabla \times \mathbf{B})$ , this integrand can be written

$$\mathbf{G} \cdot \nabla \times \nabla \times \mathbf{E} = \nabla \times \mathbf{E} \cdot \nabla \times \mathbf{G} - \nabla \mathbf{G} \times \nabla \times \mathbf{E} \quad (3.2)$$

Integrating and applying the divergence theorem to the second term gives

$$\int_{\Omega} \mathbf{G} \cdot \nabla \times \nabla \times \mathbf{E} d\Omega = \int_{\Omega} \nabla \times \mathbf{E} \cdot \nabla \times \mathbf{G} d\Omega + \int_{\Sigma} \mathbf{G} \cdot \mathbf{n} \times \nabla \times \mathbf{E} d\Sigma \quad (3.3)$$

In the surface integral,  $\mathbf{n}$  is the outward unit normal to  $\Sigma$  and the integrand has been rearranged according to the rule for scalar triple products. This identity is the vector analog of Green's identity, e.g., see Stratton, §4.14, and is simply the result of multi-dimensional integration by parts. Substituting (3.3) into (3.1), the volume-averaged scalar equation becomes

$$\int_{\Omega} \mathbf{G} \cdot (\varepsilon \ddot{\mathbf{E}} + \sigma \dot{\mathbf{E}}) d\Omega \equiv -\frac{1}{\mu_{\Omega}} \int_{\Omega} \nabla \times \mathbf{E} \cdot \nabla \times \mathbf{G} d\Omega - \frac{1}{\mu_{\Sigma}} \int_{\Sigma} \mathbf{G} \cdot \mathbf{n} \times \nabla \times \mathbf{E} d\Sigma \quad (3.4)$$

The critical result expressed in (3.4) is that the volume integral of the second order spatial operator has been replaced by "weaker" volume and surface integrals of first order operators.

#### §4.0 Reduction to an Ordinary Differential Equation

The basis for transforming the volumetric partial differential equation to an ordinary differential equation is an assumption on the mathematical form of wave fields in domain  $\Omega$ . In particular, fields are assumed to be separable in space and time, namely,

$$\mathbf{E}(\mathbf{x}, t) \equiv S(\mathbf{x})\mathbf{f}(t) \quad , \quad \mathbf{G}(\mathbf{x}, t) \equiv S(\mathbf{x})\mathbf{g}(t) \quad (4.1)$$

where  $\mathbf{x}$  is the position vector, matrix  $S(\mathbf{x})$  represents the field's spatial variation, and vector  $\mathbf{f}(t)$  or  $\mathbf{g}(t)$  represents the time variation. Note, that the same spatial variation is assumed for  $\mathbf{E}$  and  $\mathbf{G}$  in (4.1). This assumption, associated with the name of Galerkin in the finite element literature, is particularly convenient because it yields a symmetric system of equations .

Separable representation (4.1) may be interpreted in a number of ways. For example,  $S(\mathbf{x})$  can be eigenvectors (mode shapes), whence  $\mathbf{f}(t)$  are the corresponding eigenvalues (frequencies) for the given domain; or  $S(\mathbf{x})$  can be a multi-dimensional Fourier series, for which  $\mathbf{f}(t)$  are the time-dependent Fourier coefficients. Alternatively, it is easy to show that (4.1) is the functional form of a multi-dimensional Taylor series. In that case,  $\mathbf{f}(t)$  is an infinite vector of all derivatives at point  $\mathbf{x}$  and  $S(\mathbf{x})$  is a corresponding infinite matrix consisting of the coefficients of these derivatives in the series, i.e., powers of the local space coordinates. Note that each of these interpretations can provide a local or global basis for analysis.

Substituting separable solutions (4.1) into the integrands in (3.4) and applying the vector differential operator to those on the right side gives

$$\begin{aligned} \mathbf{G} \cdot \varepsilon \ddot{\mathbf{E}} &\equiv \mathbf{g}^T S^T \varepsilon \mathcal{S} \ddot{\mathbf{f}} \quad , \quad \mathbf{G} \cdot \sigma \dot{\mathbf{E}} \equiv \mathbf{g}^T S^T \sigma \mathcal{S} \dot{\mathbf{f}} \\ \nabla \times \mathbf{E} \cdot \nabla \times \mathbf{G} &\equiv \mathbf{g}^T (\nabla \times S)^T (\nabla \times S) \mathbf{f} \\ \mathbf{G} \cdot \mathbf{n} \times (\nabla \times \mathbf{E}) &\equiv \mathbf{g}^T S^T \mathbf{n} \times (\nabla \times S) \mathbf{f} \end{aligned} \quad (4.2)$$

Therefore, substituting into (3.4), moving the vector functions of time outside the integrals, and rearranging yields

$$\mathbf{g}^T \{M\ddot{\mathbf{f}} + C\dot{\mathbf{f}} - K\mathbf{f} - B\mathbf{f}\} = 0 \quad (4.3)$$

where

$$M \equiv \int_{\Omega} S^T \epsilon S d\Omega, \quad C \equiv \int_{\Omega} S^T \sigma S d\Omega, \quad K \equiv \frac{1}{\mu} \int_{\Omega} (\nabla \times S)^T (\nabla \times S) d\Omega \quad (4.4)$$

are symmetric coefficient matrices defined by the volume integrals and

$$B \equiv \frac{1}{\mu} \int_{\Sigma} S^T \mathbf{n} \times \nabla \times S d\Sigma \quad (4.5)$$

is the matrix defined by the surface integral. Since  $\mathbf{g}(t)$  is arbitrary, (4.3) is satisfied when

$$M\ddot{\mathbf{f}} + C\dot{\mathbf{f}} = (K + B)\mathbf{f} \quad (4.6)$$

This is the global ordinary differential equation equivalent to Maxwell's partial differential equations in  $\Omega$ . Of course, the utility of (4.6) depends on the choice of separable field representation, i.e.,  $S(\mathbf{x})$  and  $\mathbf{f}(t)$ .

## §5.0 The Finite Element Equations

Given the above mathematical preamble, the finite element procedure consists of partitioning or discretizing interior domain  $\Omega$  into a number of subdomains or finite elements. The field is approximated over each element by an interpolating or shape function depending on values at discrete nodes on or in the element. This yields a convenient local basis (in contrast to a global basis) for evaluating the model's matrix coefficients in (4.4) using an element-by-element summation.

To provide some degree of field continuity across element boundaries, most of the discrete nodes are defined on the element surfaces and shared by adjacent elements. Provided that they cover the domain, the elements and shape functions may be completely arbitrary. However, it is advantageous in terms of modeling and computation to make elements as simple as possible. Thus, in three-dimensions, simple hexahedron or brick shapes, i.e., so-called cartesian elements, are favored with low-degree shape functions based on nodes at the corners, and on the faces and edges as the degree of interpolant is increased. The tri-linear shape function (linear in each direction) with corner nodes is the lowest degree that provides field continuity in 3-D, hence, this is the most "elemental" interpolant. A simple analog to this shape function is the common trapezoidal rule used in numerical integration of functions in one or more dimensions.

With the domain covered by an assemblage of elements and nodes, individual elements,  $m = 1, M$ , and nodes,  $n = 1, N$ , are consecutively numbered in some convenient fashion. The global unknown vector,  $\mathbf{f}(t)$  in (4.1), is written as  $\mathbf{f}(t) = [\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3]^T$ , where vectors  $\mathbf{f}_k, k=1,2,3$  are the three components of the electric field at the  $N$  ordered nodes of the assemblage.

We consider a single, eight node, cartesian finite element or brick with element numbering and coordinates system illustrated in Fig. 1a. This is the generic element composing any assemblage, with element node numbers related to global node numbers by a simple map. For example, a simple assemblage with global numbering is shown in Fig. 1b. More will be said of this assemblage later. The shape function matrix and node vector for element  $m$  are written as

$$\mathbf{S}^m(\mathbf{x}) \equiv \begin{bmatrix} s^m(\mathbf{x}) & 0 & 0 \\ 0 & s^m(\mathbf{x}) & 0 \\ 0 & 0 & s^m(\mathbf{x}) \end{bmatrix}, \quad \mathbf{f}^m(t) \equiv \begin{bmatrix} f_1^m \\ f_2^m \\ f_3^m \end{bmatrix} \quad (5.1)$$

where  $s^m(\mathbf{x})$  and  $\mathbf{0}$  are row 8-vectors and  $f_k^m(t)$  are column 8-vectors for the three field components. The same spatial basis, i.e.,  $s^m(\mathbf{x})$ , is assumed for each field component. Note that the curl term in (3.3.4) becomes

$$\nabla \times \mathbf{S}^m(\mathbf{x}) \equiv \begin{bmatrix} 0 & -\frac{\partial s^m}{\partial z} & \frac{\partial s^m}{\partial y} \\ \frac{\partial s^m}{\partial z} & 0 & -\frac{\partial s^m}{\partial x} \\ -\frac{\partial s^m}{\partial y} & \frac{\partial s^m}{\partial x} & 0 \end{bmatrix} \quad (5.2)$$

The canonical tri-linear shape function assumed here is the row vector

$$\mathbf{s}^m \equiv \frac{1}{8} \begin{bmatrix} (1 - 2x/\Delta x)(1 - 2y/\Delta y)(1 - 2z/\Delta z) \\ (1 + 2x/\Delta x)(1 - 2y/\Delta y)(1 - 2z/\Delta z) \\ (1 + 2x/\Delta x)(1 + 2y/\Delta y)(1 - 2z/\Delta z) \\ (1 - 2x/\Delta x)(1 + 2y/\Delta y)(1 - 2z/\Delta z) \\ (1 - 2x/\Delta x)(1 - 2y/\Delta y)(1 + 2z/\Delta z) \\ (1 + 2x/\Delta x)(1 - 2y/\Delta y)(1 + 2z/\Delta z) \\ (1 + 2x/\Delta x)(1 + 2y/\Delta y)(1 + 2z/\Delta z) \\ (1 - 2x/\Delta x)(1 + 2y/\Delta y)(1 + 2z/\Delta z) \end{bmatrix}^T \quad (5.3)$$

Each component (1 to 8) of this vector is unity at the corresponding node (1 to 8) and decreases linearly to zero at all other nodes.

Given definite forms of the element shape function and nodal vector  $\mathbf{f}$ , the elemental matrix coefficients can be computed. These are found by substituting the above into integral definitions (4.4) and evaluating. For most modeling purposes,  $\varepsilon$  and  $\sigma$  may be assumed constant over an element, hence, they are typically factored out of the integration. There is no need to evaluate surface integral term  $B^m$  since it is irrelevant at the element level, although, as a point of interest it is numerically equal to  $-K^m$  by virtue of (3.2.3).

The algebra and integration of the 24x24 matrices are somewhat tedious and best done symbolically using a program like Macsyma or Mathematica. All of the evaluations described in the following sections were done in Mathematica, Wolfram (1988), because of its general utility and availability on personal computers. For example, the complete

Mathematica program to symbolically evaluate the element matrices is

```

hx = delx/2;  hy = dely/2;  hz = delz/2
s = 1/8{      (1 - x/hx) (1 - y/hy) (1 - z/hz), (1 + x/hx) (1 - y/hy) (1 - z/hz),
              (1 + x/hx) (1 + y/hy) (1 - z/hz), (1 - x/hx) (1 + y/hy) (1 - z/hz),
              (1 - x/hx) (1 - y/hy) (1 + z/hz), (1 + x/hx) (1 - y/hy) (1 + z/hz),
              (1 + x/hx) (1 + y/hy) (1 + z/hz), (1 - x/hx) (1 + y/hy) (1 + z/hz) }
dsdx = D[s,x];  dsdy = D[s,y];  dsdz = D[s,z];  zero = {0,0,0,0,0,0,0}
S = {Join[ s, zero, zero], Join[ zero, s, zero], Join[ zero, zero, s]}
curl_S = {Join[ zero,-dsdz, dsdy], Join[ dsdz, zero,-dsdx], Join[-dsdy, dsdx, zero] }
K = Integrate[ Transpose[curl_S].curl_S, {x,-hx,hx}, {y,-hy,hy}, {z,-hz,hz}]
M = Integrate[ Transpose[S].S, {x,-hx,hx}, {y,-hy,hy}, {z,-hz,hz}]

```

Dividing the resulting  $K$  by  $\mu$  and multiplying  $M$  by  $\epsilon$  gives the element matrices as a function of  $\Delta x$ ,  $\Delta y$ , and  $\Delta z$  (delx, dely, delz in the above program).

Rather than listing the 24x24 coefficient matrices, suffice it to say that they are fully populated, symmetric, and real, with  $M^m/\epsilon = C^m/\sigma$  when  $\epsilon$  and  $\sigma$  are constant. With the element coefficient matrices thus evaluated, the global system of finite element equations is assembled by inflating the element matrices, i.e., mapping the element row and column positions to the global row and column positions, and summing the contribution from each element in the assemblage.

## §6.0 Dispersion Analysis of the Finite Element Equations

The numerical solution of partial differential equations is not as straightforward as the above derivations would suggest. There are many variations on the basic numerical theme that need to be considered in the context of the physical application. An effective means of assessing pros and cons is dispersion analysis. It is particularly useful for quantifying wave-analytic properties of the numerical solutions.

Since any finite element discretization introduces an artificial length scale (element size), wave propagation through the model is necessarily dispersive, i.e., phase velocity depends on frequency. Also, properties exhibit directional dependence by virtue of the element's shape, so phase velocity is also anisotropic. This dispersive and anisotropic behavior is unavoidable in any discrete numerical solution. Fortunately, errors can be made negligible for practical purposes by insuring that element size is small compared to the minimum wavelength in the propagating signal.

The model we will use to quantify these errors consists of the 2x2x2 assemblage or molecule of cartesian elements, shown previously in Fig. 1b for the case of square elements. Since nodes are only coupled to their nearest neighbor, this simple molecule is sufficient to write the complete set of equations governing the electric field at the interior node. These equations provide the dispersion relations that completely describe wave-analytic properties of an infinite cartesian grid.

To assemble the molecule, a left-to-right, top-to-bottom, front-to-back numbering convention is used. In Fig. 1b, the three layers or sheets of nodes and the two sheets of elements are so numbered, starting from the upper left corner. The node numbering se-

quence for a single element in Fig. 1a is mapped to the molecule node numbering for each element using a simple lookup table. Inflating the element equations via this mapping and assembling them by summation yields 81x81 M and K matrices in the finite element model,  $M\ddot{\mathbf{f}} - K\mathbf{f} = 0$ . Unknown vector  $\mathbf{f}(t)$  is composed of 27 x-field, 27 y-field, and 27 z-field unknowns at the 27 nodes of the molecule. Equations for the center node are given by rows 14, 41, and 68.

We assume that a time-harmonic, linearly polarized, plane wave propagates through the model in the direction of unit wave vector  $\mathbf{k} = \{\sin\theta \cos\phi, \sin\theta \sin\phi, \cos\theta\}^T$ , specified by spherical angles,  $\theta$  and  $\phi$ . This is given by

$$\mathbf{E} = \mathbf{A} \sin\{\omega(\mathbf{k} \cdot \mathbf{x} / v - t)\} \quad (6.1)$$

where  $\mathbf{A}$  is the unit polarization vector,  $\mathbf{x}$  is the position vector, and  $v$  is the phase velocity. Given this prescription of the incident field vector,  $\mathbf{f}(t)$  is found directly by evaluating (6.1) at the  $x$  coordinate of each of the 27 nodes. Instead of simply differentiating (6.1) to obtain  $\ddot{\mathbf{f}}$ , it is evaluated from the central difference approximation

$$\ddot{\mathbf{E}}(t) \approx (\mathbf{E}(t + \Delta t) - 2\mathbf{E}(t) + \mathbf{E}(t - \Delta t)) / \Delta t^2 \quad (6.2)$$

corresponding to the typical step-wise forward integration scheme with timestep  $\Delta t$  and error on the order of  $\Delta t^4$ .

The problem is to determine phase velocity  $v$  compatible with the finite element equations and assumed form of the propagating field. This is accomplished by evaluating solution vectors  $\mathbf{f}$  and  $\ddot{\mathbf{f}}$  at the nodes from (6.1,2), isolating the three field equations for the interior node in  $M\ddot{\mathbf{f}} - K\mathbf{f} = 0$ , i.e., rows 14, 41, and 68, and solving for  $v$  and  $\mathbf{A}$ . It is convenient to rewrite the three equations as  $V\mathbf{A} = 0$  where  $V$  is a 3x3 matrix coefficient. This system of equations has a solution if and only if the determinant of  $V$  vanishes, i.e.,  $|V| = 0$ , yielding a nonlinear scalar equation on phase velocity  $v$ . This is the dispersion relation and its solution near  $c = 1/\sqrt{\epsilon\mu}$  must be found numerically. Substituting  $v$  back into  $V$  gives the homogeneous system of equations governing polarization vector  $\mathbf{A}$ . In particular,  $\mathbf{A}$  is the space mapped by linear transformation  $V$  into the null vector, i.e.,  $\mathbf{A}$  is equal to the null space of  $V$ . These solutions are described below.

Near  $v = c$  the dispersion relation is found to be second order in general, i.e., it looks like a parabola locally. Therefore it exhibits two solutions near  $c$ . This multiplicity can be traced to the element shape function row vector, (5.3). Evaluating the products yields terms proportional to  $1, x, y, z, xy, xz, yz, xyz$  for each vector component, corresponding to the first 8 terms in a Taylor series (recall, there are 8 nodal values, hence, 8 terms in the series are determinate). However, by a simple rotation of coordinates ( $X, Y, Z$  say) the last four terms may be converted to product terms like  $X^2, Y^2, Z^2$ , and  $X^3, Y^3, Z^3$ . Therefore, the shape function's form is not rotationally invariant, unlike the original differential equation. This "rotational variability" shows itself as two phase velocities in each direction.

Numerical dispersion results for the case of a cubic cartesian grid are plotted in Figure

3, showing  $v/c$  versus spherical angles of incidence in Fig. 3a and maximum  $v/c$  versus discretization in Figure. 3b. Definition of the spherical angles is shown in Figure 2. Discretization is measured by the number of elements supporting a wavelength. Shape of the normalized  $v$  surface in Fig. 3a is nearly independent of discretization so only one plot is shown, for 20 elements/wavelength in this case. Note that maximum error occurs along the cartesian axes and phase velocity is always greater than  $c$ .

Eigenanalysis of  $V$  shows that, in general, this 3x3 matrix is doubly degenerate, i.e., there is one nonzero eigenvalue and two “zero” eigenvalues that are clustered very near zero. Eigenvectors of these “zero” eigenvalues span the nullspace of  $V$ —a plane in this case. Vector  $A$  lies in this so-called polarization plane, which is perpendicular to the eigenvector of the nonzero eigenvalue of  $V$ , denoted  $q$ . Since Maxwell’s equations represent transverse waves, ideally the polarization plane is perpendicular to wave vector  $k$ , i.e.,  $q$  and  $k$  are colinear, however, grid anisotropy produces a “transversality” error. This is measured by the angle between  $q$  and  $k$  and is plotted in Fig. 4 for the cubic cartesian grid. Figure 4a shows the normalized transversality error versus spherical angles and Fig. 4b shows its maximum value versus discretization. As with the phase velocity, shape of the transversality error surface is only very weakly dependent on discretization, hence, one plot suffices.

## §7.0 Approximate Finite Element Equations

Dispersion analysis of the exact finite element equations indicates four difficulties. The first is philosophical, namely, that the numerical phase velocity in vacuum is greater than the continuum speed of light—which physical objects can never exceed. It would be preferable to have a conservative solution where numerical velocities are always less than continuum velocities.

The second difficulty is that phase velocity errors are greatest in the local coordinate directions, i.e., normal to the element faces. In practice it is natural to align numerical models with these coordinates and also to gather information on waves traveling along or near them, e.g., for imaging. Therefore, it would be better if propagation errors were minimized rather than maximized in these directions.

The third difficulty is existence of two waves in any direction. This is typical of anisotropic media, where the two are denoted ordinary and extraordinary waves. Since the phase velocities are close, the two waves are numerically indistinguishable in most cases. Nonetheless, from an analytical viewpoint, particularly with respect to boundary conditions, this is an unwelcome complication.

The fourth difficulty is the number of floating point operations necessary to evaluate and solve the exact finite element equations. For example, there are at least a factor of 5 more than necessary for a conventional finite difference approximation of Maxwell’s equations. This profligacy is a deterrent to finite elements despite their marked advantage for

modeling geometrically complicated features.

These difficulties are known to numerical wave propagation analysts, particularly in the elasticity community. One approach that claims to remedy all of them is approximation of the coefficient matrices by parameter lumping and reduced integration. Parameter lumping diagonalize the  $M$  and  $C$  matrices by placing the sum of each row in the diagonal position and zeroing the off-diagonal terms. Reduced integration applies to evaluation of the  $K$  matrix, using an approximate quadrature rule, e.g., the simple rectangular rule, or more generally, single-point Gaussian quadrature.

Dispersion analysis of the lumped parameter, reduced integration finite element equations is done in the same way as for the exact equations. Corresponding results are plotted in Figures 5 and 6. The dispersion relation is still locally parabolic but just reaches zero at its maximum (or minimum depending on the sign chosen) so there is only one phase velocity. Comparing Figs. 3 and 5, velocity dependence on angle of incidence is qualitatively similar but numerical phase velocities are always less than exact phase velocities. Maximum error occurs near the space diagonal ( $\theta = 45^\circ$ ,  $\phi = 45^\circ$ ) and Fig. 5b shows that absolute phase velocity error is only slightly greater than that exhibited by the exact equations.

Clearly, the approximate finite element equations solve the first three problems mentioned at the beginning of this section. A count of arithmetic operations shows that they also require less than half the operations needed to evaluate the exact equations. The physical basis for these approximations can be found in the differential and integral forms of Maxwell's equations, (2.1,2). In particular, the lumped parameter and reduced integration finite element equations can be shown to follow from the first of (2.1) and the second of (2.2) written as

$$\dot{\mathbf{H}} = -\frac{1}{\mu} \nabla \times \mathbf{E} \quad , \quad \int_{\Omega} \epsilon \dot{\mathbf{E}} d\Omega = \int_{\Sigma} \mathbf{n} \times \mathbf{H} d\Sigma \quad (7.1)$$

for the case of zero conductivity.

## §8.0 Radiation Boundary Conditions

Discrete numerical methods like finite elements or finite differences are necessarily formulated on finite spatial grids—whether the actual domain being modeled is finite or infinite. This domain truncation introduces artificial boundaries that must be treated with special care in order to minimize nonphysical wave reflections. These trap energy that would otherwise be radiated and establish undesirable resonances within the grid. Note that this is true regardless of the solution scheme applied, in either the time- or frequency-domain. The solution to the problem is to apply so-called radiation or absorbing boundary conditions on the model's exterior surfaces.

In optics-type problems there is an added complication because some form of illumination is usually prescribed over the model. This is accommodated by an illumination boundary condition, designed to apply the known incident electromagnetic field on the model boundaries and transmit or absorb the scattered field as if the model extended out to infinity. Determination of an incident electric field consistent with the numerical propagation characteristics of the grid is a non-trivial calculation, especially in the presence of arbitrary angles of incidence and model topography. These numerical problems are addressed elsewhere in this report. Here it is assumed that the incident field is known, and approximate radiation conditions for the scattered fields are described.

In formulating radiation boundary conditions, there is a tradeoff between accuracy and complexity. Generally, the more boundary nodes that are coupled by the boundary formulation, the more accurate and computationally intensive the condition will be for a given model size. Increasingly accurate radiation conditions can, in principle, allow smaller and smaller models around the scattering features of interest. Development along these lines may well be warranted, but for this report attention is restricted to absorbing conditions that are local in space and time for compatibility with the explicit time integration approach. In this way no more than a small fraction of the computational effort is expended on the evaluation of boundary conditions.

Here, the basis for radiation conditions is the paraxial wave equation, i.e., an equation valid for propagation in (and around) a selected direction. The prototypical example is the 1D wave operator, which can be factored as

$$\left\{ \frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} \right\} E = \left\{ \frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right\} \left\{ \frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right\} E = 0 \quad (8.1)$$

where the factors

$$\left\{ \frac{\partial}{\partial t} \pm c \frac{\partial}{\partial x} \right\} E = 0 \quad (8.2)$$

represent the right and left traveling waves. Thus, if the time and space derivatives are related according to this operator at the ends of a 1D domain then an “exact” radiation boundary condition results; it cannot be truly exact in practice because approximations are implicit in the numerical implementation. This is the well-known normal incidence condition, the so-called “black” boundary. It turns out to be quite reasonable for many applications, particularly for boundaries very far from a compact body in a homogeneous material, e.g., for conventional radar-like scattering problems.

In order to significantly enhance boundary condition performance, higher order paraxial approximations of the multidimensional Maxwell’s equations must be derived and applied as discussed below. Specifically, the 4<sup>th</sup> order condition of Clayton and Enquist is presented for Maxwell’s equations in 2, 2.5 and 3 space dimensions. Also, a condition proposed by Sandler [1991] is described. The 4<sup>th</sup> order paraxial and Sandler conditions are demonstrated to be roughly equivalent in accuracy, although certain implementation issues appear to favor the latter.

### Paraxial Absorber

In their paper, “Absorbing Boundary Conditions For Acoustic And Elastic Wave Equations,” [1977], Clayton and Engquist developed an absorbing boundary condition for the 2D elastic wave equation using a paraxial approximation method. This approach has been extended here to Maxwell’s equations with zero conductivity. For the sake of brevity only the resulting paraxial equations are stated, since the derivation is not immediately relevant. Engquist [1991] claims that extension to the case of finite conductivity should also be possible.

The paraxial versions of Maxwell’s equations are derived from the equivalent second order partial differential equation on  $E$ , (2.5). In the following equations it is assumed that the x-coordinate is normal to the absorbing surface. In 2D, the paraxial equation is,

$$E_{tt} \pm cE_{tx} - \frac{c^2}{2}(E_{yy}) = 0 \quad (8.3)$$

where  $c = 1/\sqrt{\epsilon\mu}$  denotes the wavespeed, while in 3-D this becomes

$$E_{tt} \pm cE_{tx} - \frac{c^2}{2}(E_{yy} + E_{zz}) = 0 \quad (8.4)$$

The paraxial equation for 2.5D is

$$E_{\tau\tau} \pm \frac{c}{\sin\theta}E_{\tau x} - \frac{c^2}{2\sin^2\theta}(E_{yy}) = 0 \quad (8.5)$$

where  $\theta$  is the angle between the wave vector and the z-axis. The Galilean transformation

$$\tau = t - \frac{\cos\theta}{c}z \quad (8.6)$$

is the basis for the 2.5D transformation that eliminates time  $t$  and axial dimension  $z$ . These are so-called 4<sup>th</sup> order paraxial forms, in contrast to (8.2), which is the 2<sup>nd</sup> order form. Order refers to the order of the approximation, effectively in terms of cone angle around the paraxial direction.

The purpose of the exercise is to derive a consistent boundary condition from these paraxial approximations of Maxwell’s equations. This is accomplished by applying the standard Galerkin finite element formulation to equations (8.3-5), from which an implementation of the boundary conditions consistent with the interior domain may be derived. However, there does not appear to be a single, consistent way of doing this, particularly in the context of the reduced integration techniques used in finite element algorithms.

For the 2D case, no difficulties are encountered. In 2.5D and 3D certain terms are observed to excite hourglass modes due to the single-point quadrature used by EMFlex. These are currently removed, resulting in a loss of potential absorber accuracy. Clearly, more work on the implementation of paraxial boundary conditions in finite element algorithms is required. It appears that a complete paraxial finite element formulation with a layer of paraxial elements around the model instead of merely paraxial boundaries

will be a useful direction for development.

#### Sandler Absorber

Although effective, the 4<sup>th</sup> order paraxial conditions are not naturally compatible with discrete algorithms, nor do they readily admit conductivity. The problem is that these methods are based on an analytical approximation of Maxwell's equations rather than on an approximation of the discrete finite element form of these equations. There is a subtle but real difference. Approaches that attack the discrete equations themselves have been studied recently. In particular, Sandler [1991] developed a mechanically based absorbing condition for nonlinear solid mechanics applications. This condition is superior to the paraxial approach described above because it eliminates wavespeed from the formulation (which in nonlinear calculations is problematic), and it provides better directionality than the 2<sup>nd</sup> order paraxial or normal incidence condition.

Because of its directionality characteristics we derived the Sandler condition for the 2D scalar wave equation (Maxwell's equations for E-perpendicular polarization) and compared it to the 4<sup>th</sup> order paraxial condition. This implementation is based on physical arguments that are tied directly to the discrete finite element equations rather than a modified partial differential equation.

#### Absorber Comparison

A small 2D test problem is used to evaluate effectiveness of the various boundary conditions described above. The computational domain for this problem is a 80 x 160 element grid. The source consists of a transient electric field applied over a 2x2 element square region towards one end of the grid. The outward travelling waves strike the upper and lower boundaries at angles varying from normal incidence to almost grazing incidence. This provides a comprehensive test of absorber capabilities. An "exact solution" was obtained from an extended grid calculation, selected such that spurious boundary reflections would not appear in the 80 x 160 element window over the duration of the test.

Figures 7-10 show a snapshot sequence of the evolving electric field for the exact, Fig. 7, normal (2<sup>nd</sup> order), Fig. 8, paraxial (4<sup>th</sup> order), Fig. 9, and Sandler, Fig. 10, absorbers. The background grid noise in the exact solution is due to a small amount of hourglassing excited by the 2x2 element square transient source region. By comparing the snapshots, it is evident that the 4<sup>th</sup> order paraxial and Sandler absorbers are roughly equivalent and that both are superior to the normal incidence condition. The Sandler condition thus provides an effective alternative to paraxial conditions, but with better "discreteness" compatibility and the ability to accommodate conductivity.

### **§9. Time Domain versus Frequency Domain**

Why does EMFlex solve steady-state problems with a time-domain algorithm? This is an obvious, fair, and often asked question. The short answer is that time-domain methods currently provide the quickest, least computer memory intensive, most robust path to a solution. Advances in numerical methods may someday change this answer. There do

exist reasons for using frequency-domain simulations if they become competitive in CPU and memory cost. In this section, we outline the issues governing the choice between time- and frequency-domain solvers for finite element models, list the pros and cons of each approach, present some comparisons between EMFlex and an equivalent frequency domain formulation for the 2-D scalar case (using state-of-the-art iterative solvers), and conclude with some observations on the future outlook for frequency-domain computations.

The primary advantage of the time-domain solver used in EMFlex is that it embodies an efficient explicit algorithm that requires minimal memory, thus permitting solutions of the largest finite element model possible on any given machine. It is also robust and deterministic in the sense that if run long enough, steady-state will be achieved and a solution will be found. Additional advantages of the time-domain approach are that it is directly extensible to nonlinear problems where material properties change with time, e.g., the bleaching process in photolithography, wave arrivals may be separated in time providing better insight into physical processes, and transient (pulsed) or other nonsinusoidal signals are easily accommodated.

Disadvantages of time-domain solvers are that user intervention is required to assess when steady-state has been reached. One can envision problems where a long simulation may be necessary to achieve steady-state. Also, the steady-state quantities (amplitude and phase) are not primary in the time domain, and must be obtained by a secondary calculation after steady-state has been achieved. Note that this process has been automated in EMFlex.

Given that we are looking for a frequency-domain solution, the advantages of a straightforward frequency-domain formulation appear obvious. The difficulty is that for realistically sized models this formulation requires the solution of an enormous, sparse, linear system of equations. In addition, the linear system has undesirable numerical properties, namely, it is complex, non-hermitian (due to absorbing boundary conditions and/or conductivity), non-symmetric (due to absorbing boundary conditions), and is typically indefinite.

There are two basic methods of solving linear systems: direct (some variation of Gaussian elimination) and iterative. Direct methods are typically divided into two phases: factorization of the coefficient matrix, followed by back-substitution to obtain the solution. The payoff here is that most of the work is performed in the factorization, so that solutions for additional right hand sides, e.g., different illumination sources, can be obtained from a relatively cheap backsolve. The drawback of direct methods is that they require large amounts of memory and CPU time. For example, consider a  $100 \times 100 \times 100$  element 3D model with 3 electric field components unknown at each node, in complex arithmetic. This is a modest model size. For this problem, a standard band solver such as ZBGFA out of LINPACK [Donagarrá et al, 1979] requires  $5.73 \times 10^{11}$  words of storage, which far exceeds the capacity of any present day computer. The number of floating point operations to solve such a system is also prohibitive. It is worth noting here that this linear

system is very sparse, i.e., only  $2.5 \times 10^8$  or 0.04% of the entries are nonzero. Observe that even the nonzero entries exceed the largest available machines, but only by a small margin. Some work has been done on sparse system direct solvers which attempt to economize on storage relative to the basic band solvers, but our experience with one such routine (MA28 from the Harwell library) was disappointing. Thus we concur with the conventional wisdom which holds that 3D problems cannot be solved directly.

The alternative to direct solvers is iterative methods in which an initial guess for the solution (zero if nothing better is available) is successively refined until the error becomes “small.” Conventional wisdom says that the Krylov subspace-type methods are best, which includes the conjugate gradient approach. Frequently, some type of preconditioner is used in conjunction with these methods to accelerate convergence. For some classes of matrices, e.g., positive definite symmetric, these methods work extremely well. The major portion of the CPU effort is in computing a product of the coefficient matrix with a vector (a “MATVEC” operation) and in computing inner products of two vectors. The MATVEC result can be obtained in an explicit element-by-element operation as is done in EMFlex for the time-domain solver. Thus, the insurmountable memory requirements of the direct methods can be avoided. Additional vectors of length  $N$  (the total number of unknowns), which form the basis of the Krylov Subspace, are usually required. For some iterative methods such as GMRES, these add up to many times more memory than for the time-domain algorithm.

Things to consider in choosing an iterative method are: 1) the total amount of work required by the method and the preconditioner to achieve an acceptable solution; 2) the total amount of memory required by the method and the preconditioner; 3) the possibility of breakdown—the algorithm fails, or hangs—the convergence rate becomes so low that a solution is not attainable; and 4) determinism—a solution might be obtainable only by trial-and-error tweaking of the preconditioner or iterative solver parameters.

### Examples

We compare two iterative solvers to EMFlex for a small 2D scalar example. Our vehicle is a frequency domain finite element code called Invmax, which was written by us under an NSF grant to investigate inversion. It is restricted to the 2D scalar case, uses the lowest order absorbing boundary condition, and employs exact integration of the element matrices. To limit the differences in the codes as much as possible, we used the same absorber in EMFlex and incorporated a special 2D processor with exact integration for this exercise.

The iterative solver used by Invmax is LSQR [Paige and Saunders, 1982]. We adopted the obvious approach of converting the ( $N \times N$ ) complex system into a ( $2N \times 2N$ ) equivalent real system, which is nonsymmetric and indefinite. Freund [1991] has recently presented an analysis which clearly demonstrates that this is not a wise thing to do and that one should solve the complex system directly. LSQR solves this system by applying the standard conjugate gradient method to the normal equations. In other words, it obtains

the solution to the nonsymmetric, indefinite system

$$\mathbf{Ax} = \mathbf{b} \tag{9.1}$$

by solving the symmetric positive definite system,

$$\mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{b} \tag{9.2}$$

While the three term recursion used in the standard conjugate gradient method uses minimal storage for an iterative method, this procedure has two disadvantages. It turns out that each LSQR iteration requires 2 MATVEC operations ( $\mathbf{A}^T \mathbf{A}$  is never computed explicitly), which is double the work per iteration of solving (9.1) directly. Also, the condition number (ratio of largest to smallest eigenvalue) of  $\mathbf{A}^T \mathbf{A}$  is the square of that for  $\mathbf{A}$ , so many more iterations are required to attain a solution.

To see if this result could be improved upon, we collaborated with a company of computer scientists (Advanced Scientific Computation) to make Invmax as efficient as possible using off-the-shelf linear solvers of their choice. The outcome of this was a new version of Invmax which they called Ascus. Ascus implements a version of GMRES [Burkhart and Young, 1988] adding optional preconditioning obtained by solving the real part of the coefficient matrix with a band-solver. This preconditioner requires about 25% of the work of solving the complex system directly and involves too much memory to do anything but small 2D problems. Their idea for extension to larger problems is to replace the band solver with an incomplete factorization of the real part. This would not be as effective as the band solver but should result in fewer iterations than for no preconditioner. An incomplete factorization still requires storing all of the non-zero entries of the coefficient matrix, which is prohibitive for 3-D problems.

The goal of this exercise is to compare the effort necessary to obtain a reasonable solution. Since the finite element approximation introduces errors on the order of a few percent, and the time- and frequency-domain formulations will produce slightly different results, we adopt the following criterion for comparison. Each method is used to generate its own accurate solution—after a large number of iterations or timesteps. For each method we define this as the “true” solution, and measure RMS error on a line above the feature versus timestep or iteration. A reduction in RMS error to 5% is deemed reasonable, and sufficiently removed from the “true” solution to be insensitive to this relatively arbitrary choice.

Because EMFlex is coded more efficiently than Invmax or Ascus, it would be misleading to compare CPU times directly. Invmax or Ascus could be brought to the EMFlex level of efficiency with an investment of programming time. Assuming that most of the work in the iterative methods is devoted to the MATVEC operation we make the following equivalence. Since the frequency-domain systems are  $2N \times 2N$ , one MATVEC is in general equivalent to 4 timesteps in EMFlex, or 2 timesteps if we know a priori that there are very few imaginary coefficients (no conductivity). We will say that 1 MATVEC is equal to 2 timesteps.

Figure 11 shows converged solutions for the example problem, which is a 1x1 micron

oxide line on a silicon substrate. The grid is  $91 \times 101$  nodes for a total of 9000 elements. An outline of the feature and the output line are shown. The scattered electric field is pictured on the left and plotted along the output line on the right. We observe that the converged solutions are very similar for the three calculations. Figure 12 shows the evolution of RMS error for each method. A 5% RMS error requires about 250 timesteps in EMFLex, 3200 equivalent timesteps in LSQR, 600 equivalent timesteps in GMRES without preconditioning, and 130 equivalent timesteps in GMRES with preconditioning. Considering the cost of the preconditioner, it is not clear whether the preconditioned or unpreconditioned GMRES is better. We somewhat arbitrarily chose 15 Krylov vectors for GMRES. Retaining more vectors would require more memory and more work per iteration, but presumably fewer iterations.

It is apparent from this example and larger ones not shown here, that the GMRES solver is far superior to the LSQR approach in CPU requirements, at the cost of increased memory. However, neither approach can equal the time domain solver in either memory or CPU time.

### Conclusions and Future Outlook

The frequency domain calculations reported above represent the best one can do with readily available state-of-the-art solvers. Recently, Freund [1991, 1992] proposed a new Krylov-subspace method which he calls QMR (for quasi-minimal residual) and extended it to general nonsingular, non-Hermitian systems. Prompted in part by our inquiries a few years ago, he specifically targets the systems arising from frequency-domain discrete numerical methods. Preliminary results appear far superior to GMRES, and QMR has the additional advantage that only a couple of vectors need to be stored. He has also devised a strategy that circumvents one cause of breakdown, though incurable breakdowns are still possible in theory. A general-purpose implementation of this method will probably be available within a year or two for evaluation.

For frequency-domain finite element solvers, we conclude that solution techniques for the resulting class of linear systems are still in the research phase. Significant advances have been achieved in recent years, and more advances are expected. We believe that trying new solution algorithms is a worthwhile exercise which provides feedback to the mathematicians and computer scientists, and is to everyone's advantage. On the other hand, it is premature to promote such solvers as production level tools for engineers. We conclude that if large-scale calculations need to be done today, time-domain techniques provide the most practical means of doing them.

## §10.0 References

- Burkhart, R.H. and D.P. Young, "GMRES Acceleration and Optimization Codes," ETA-TR-88, Boeing Computer Services, May 1988.
- Burkhart, R.H. and D.P. Young, "Documentation for GMRES Acceleration and Optimization Codes," ETA-TR-89, Boeing Computer Services, May 1988.
- Clayton, R. and Björn Engquist. "Absorbing boundary conditions for acoustic and elastic wave equation," *Bull. Seism. Soc. Amer.* **67**, 1977, pp. 1529-1540.
- Dongarra, J.J., C.B. Moler, J.R. Bunch and G.W. Stewart, 'LINPACK User's Guide,' SIAM, 1979.
- Engquist, B. Personal Communication, Dept. of Mathematics, University of California, Los Angeles, 1991.
- Freund, R.W., "Krylov Subspace Methods for Complex Non-Hermitian Linear Systems," Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffett Field, Ca., 1991.
- Freund, R.W., "Conjugate Gradient Type Methods for Linear Systems with Complex Symmetric Coefficient Matrices," *SIAM J. Sci. Stat. Computing*, Vol.13, No. 1, Jan. 1992 (Preprint).
- Paige, C.C., and M.A. Saunders, "LSQR: An algorithm for sparse linear equations and sparse least squares," *ACM Transactions on Mathematical Software*, March 1982.
- Sandler, I., Personal Communication, Weidlinger Associates, New York, NY, 1991.
- Stratton, J. A. (1941). *Electromagnetic Theory*, McGraw-Hill Book Co., New York, NY.
- Papas, C. H. (1988). *Theory of Electromagnetic Wave Propagation*, Dover Publications, Inc., New York.
- Wolfram, S. (1988). *Mathematica: A System for Doing Mathematics by Computer*, Addison-Wesley Publishing Co., Inc.
- Zienkiewicz, O. C. (1977). *The Finite Element Method*, 3<sup>rd</sup> Edition, McGraw-Hill Book Co. (UK) Limited.

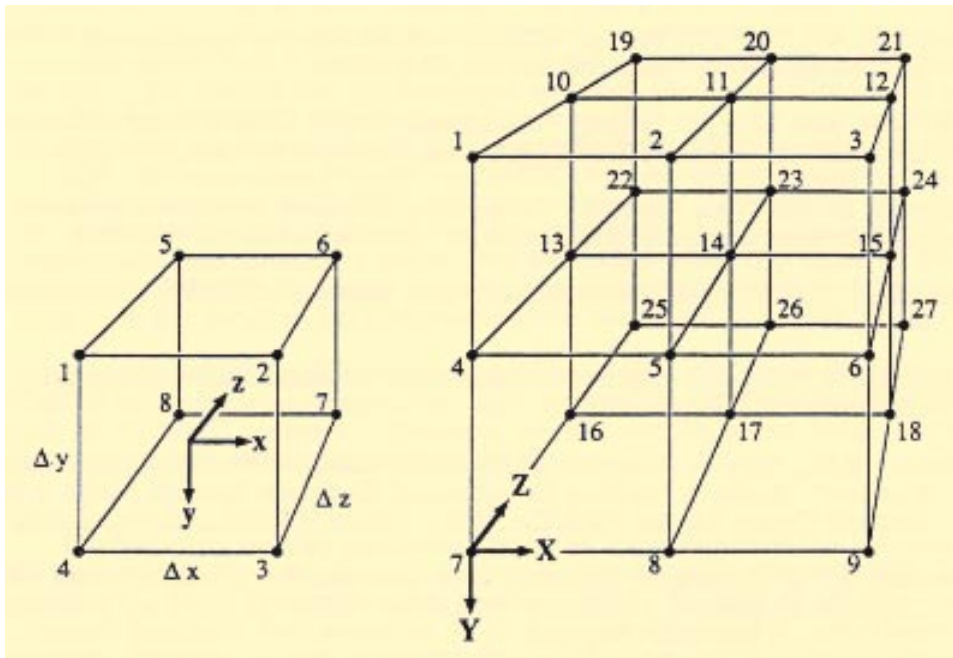


Figure 1a

Figure 1b

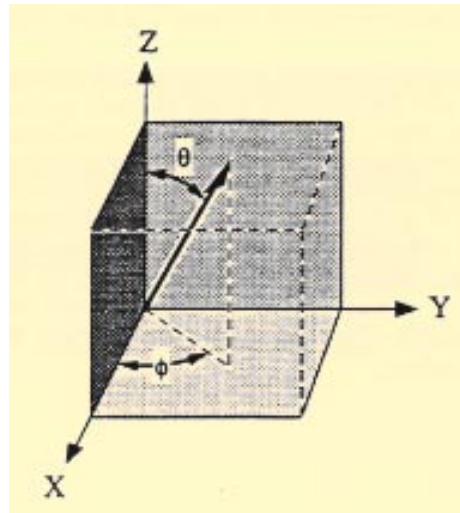


Figure 2

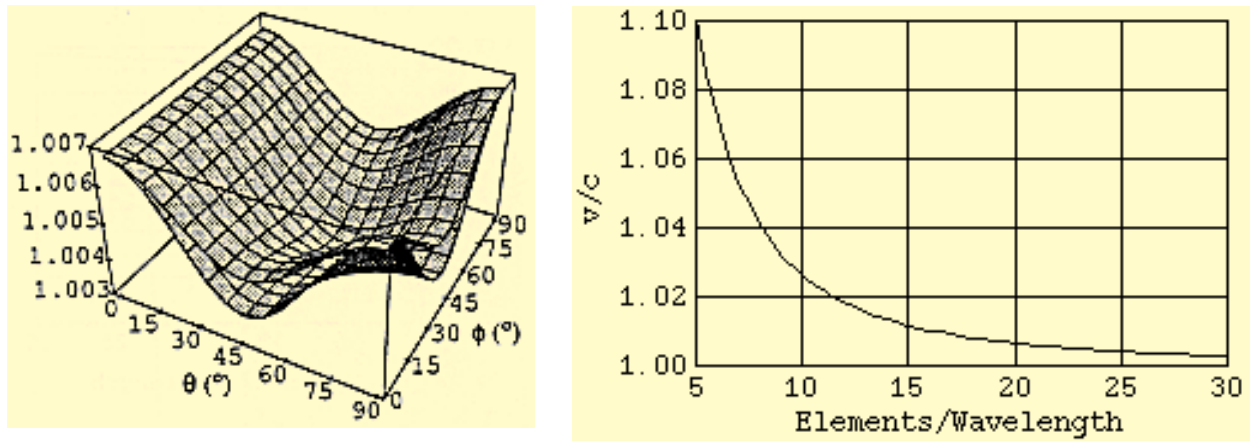


Figure 3. Results from dispersion analysis of the exact finite element equations for a cubic element grid. Phase velocity versus spherical angles of incidence is shown on the left for the case of 20 elements/wavelength. The two surfaces are the two velocities satisfying the dispersion relation. Maximum phase velocity versus discretization is plotted on the right.

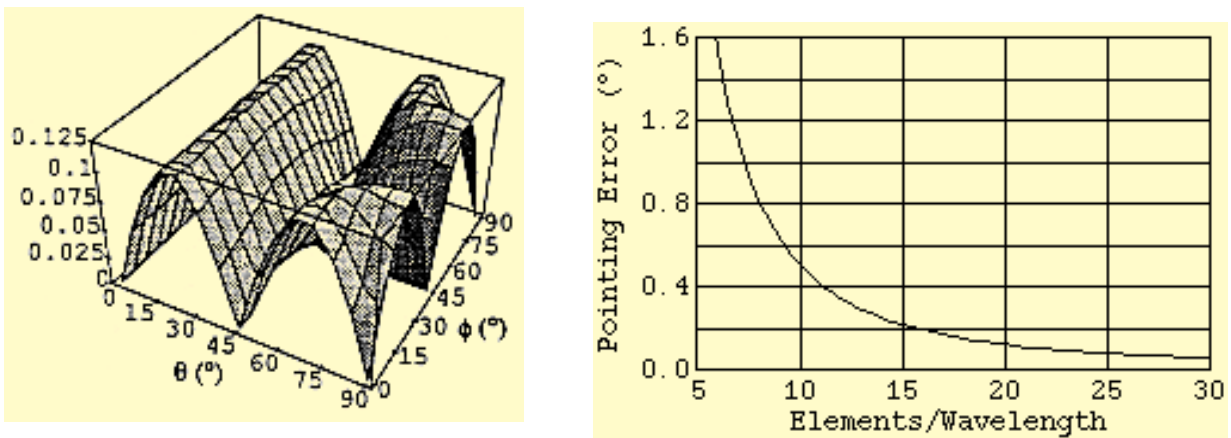


Figure 4. Transversality or pointing error versus spherical angles of incidence in the finite element grid of Figure 3 is shown on the left for the case of 20 elements/wavelength. Absolute pointing error is virtually identical for the two velocities satisfying the dispersion relation shown in Figure 3. Maximum pointing error versus discretization is plotted on the right.

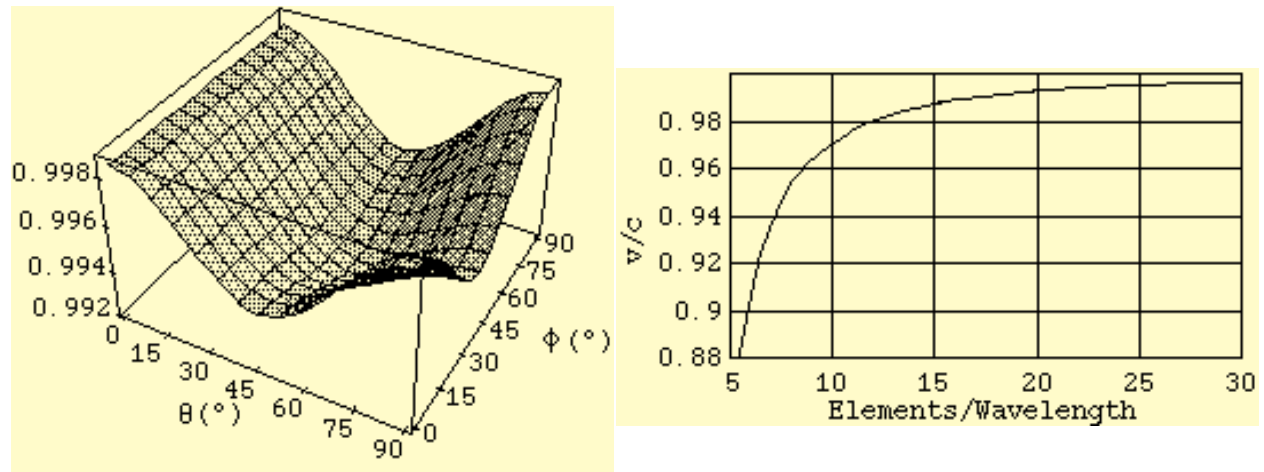


Figure 5. Results from dispersion analysis of the approximate finite element equations for a cubic element grid. Phase velocity versus spherical angles of incidence is shown on the left for the case of 20 elements/wavelength. Minimum phase velocity versus discretization is plotted on the right.

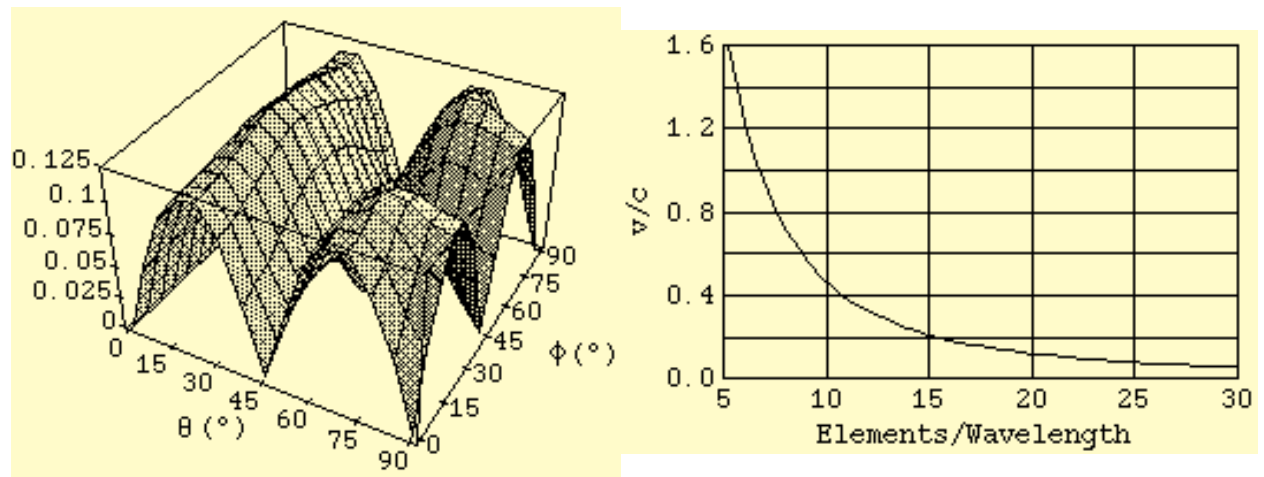


Figure 6. Transversality or pointing error versus spherical angles of incidence in the finite element grid of Figure 5 is shown on the left for the case of 20 elements/wavelength. Maximum pointing error versus discretization is plotted on the right.

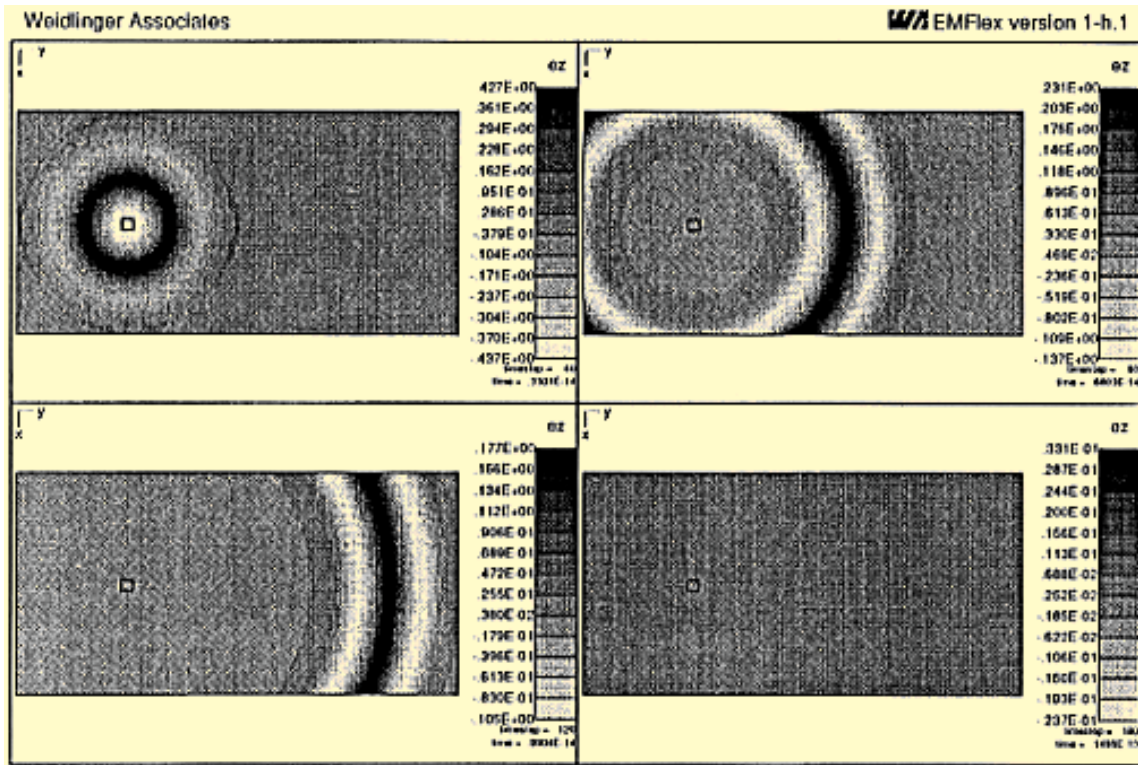


Figure 7. Snapshot sequence of the exact finite element solution within the nominal grid (by plotting the interior of an extended grid).

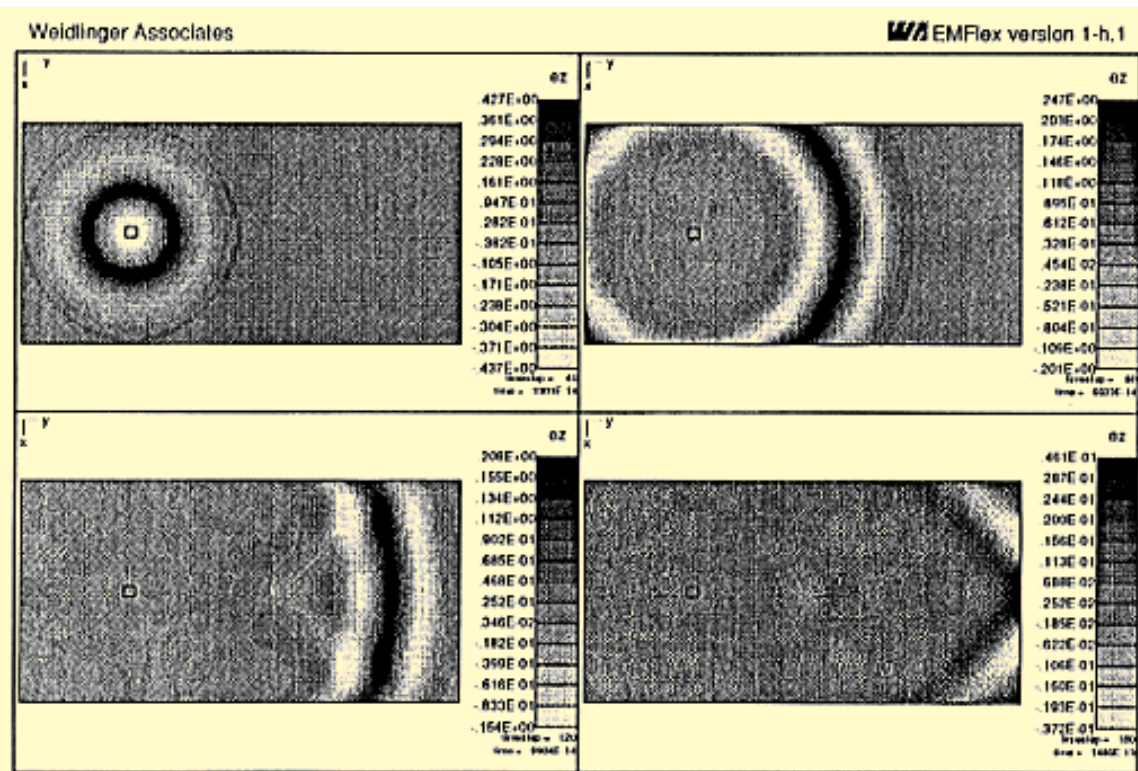


Figure 8. Snapshot sequence of the normal incidence (2nd order paraxial) boundary condition on the model's exterior.

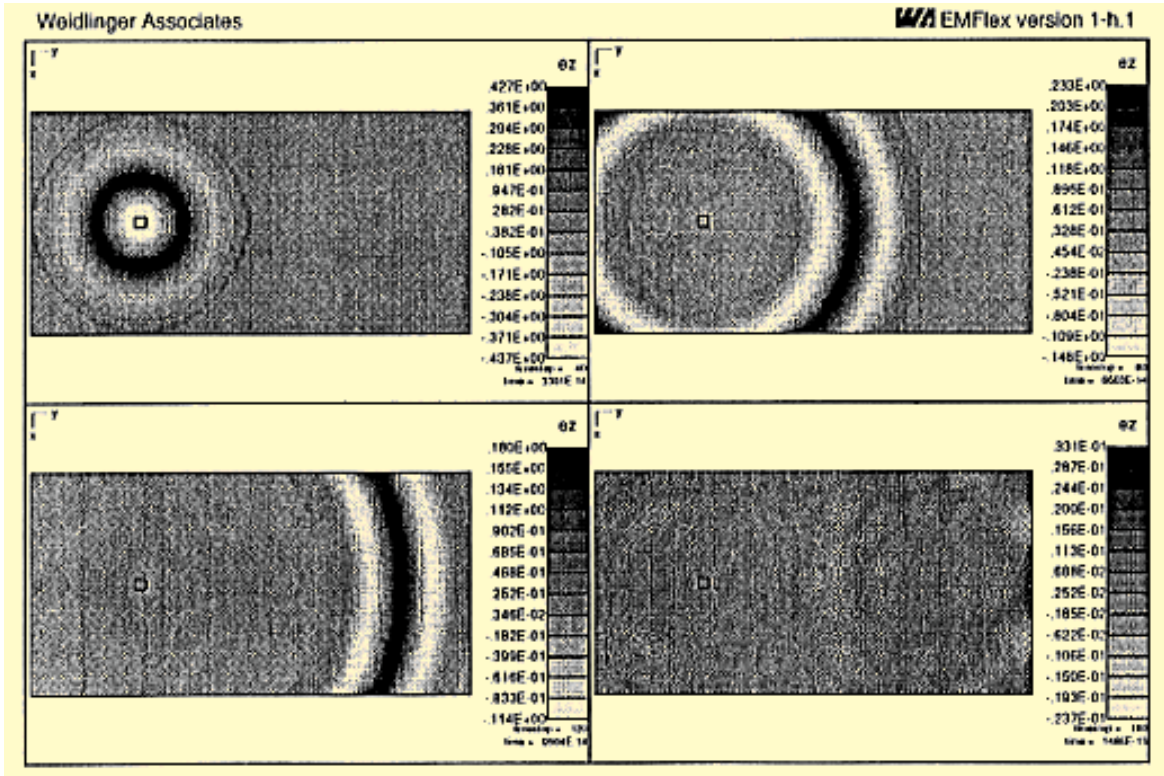


Figure 9. Snapshot sequence for the paraxial (4th order paraxial) boundary condition on the model's exterior.

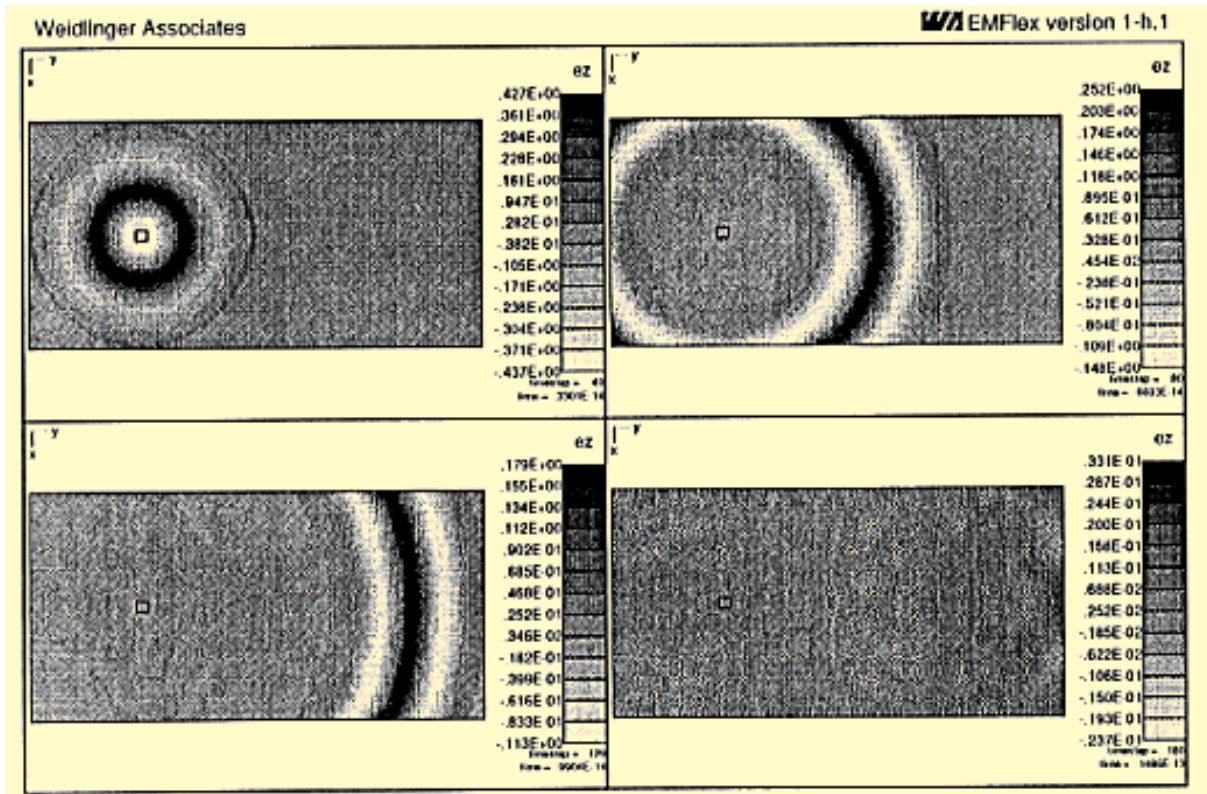


Figure 10. Snapshot sequence for the Sandler boundary condition on the model's exterior.

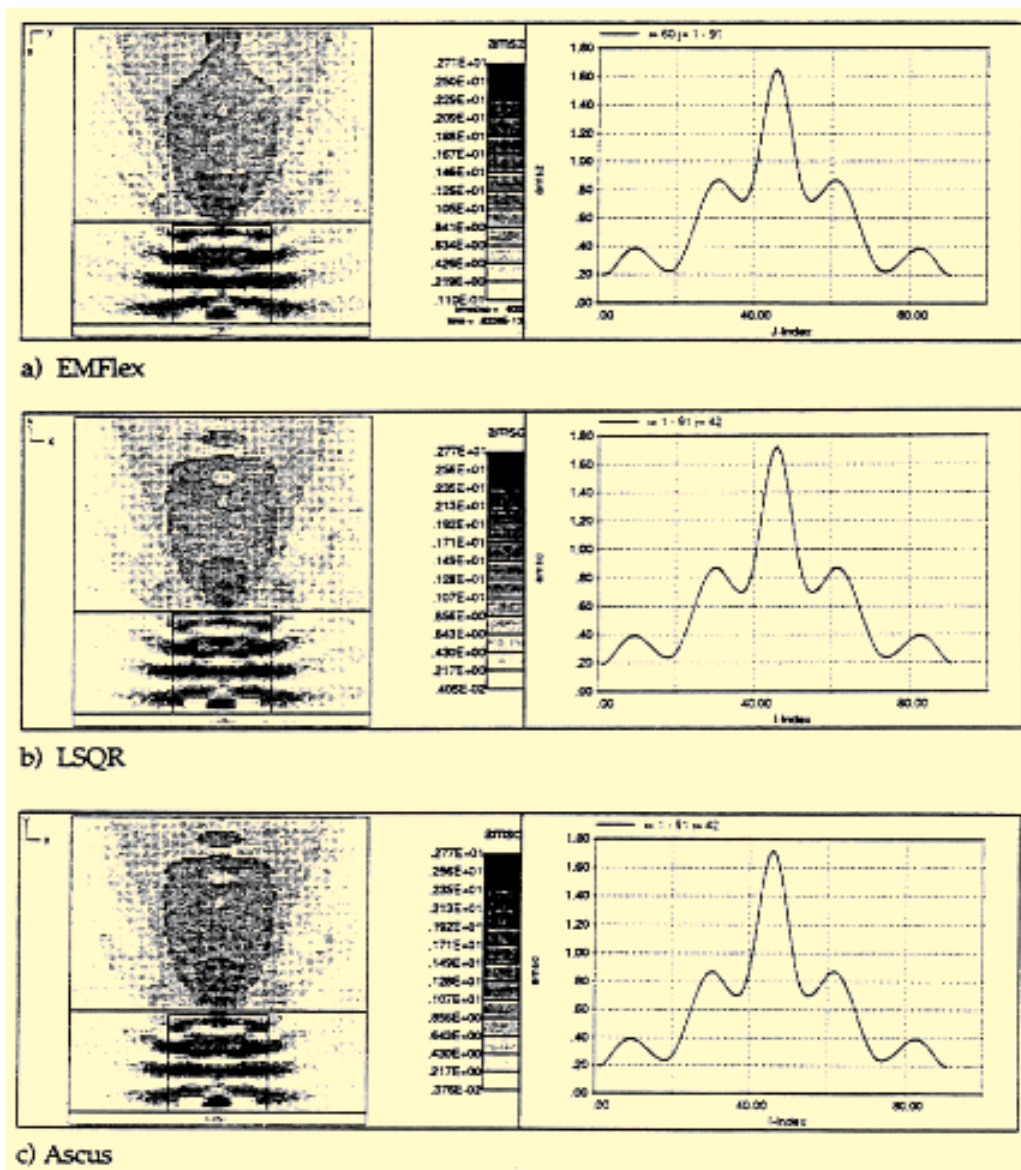


Figure 11. Comparison of converged solution for the time- and frequency-domain algorithms. The 9000 element model represents a  $1.0 \times 1.0$  micron oxide line on bare silicon. Snapshots on the left show the field distribution and plots on the right give the field on a line just above the feature.

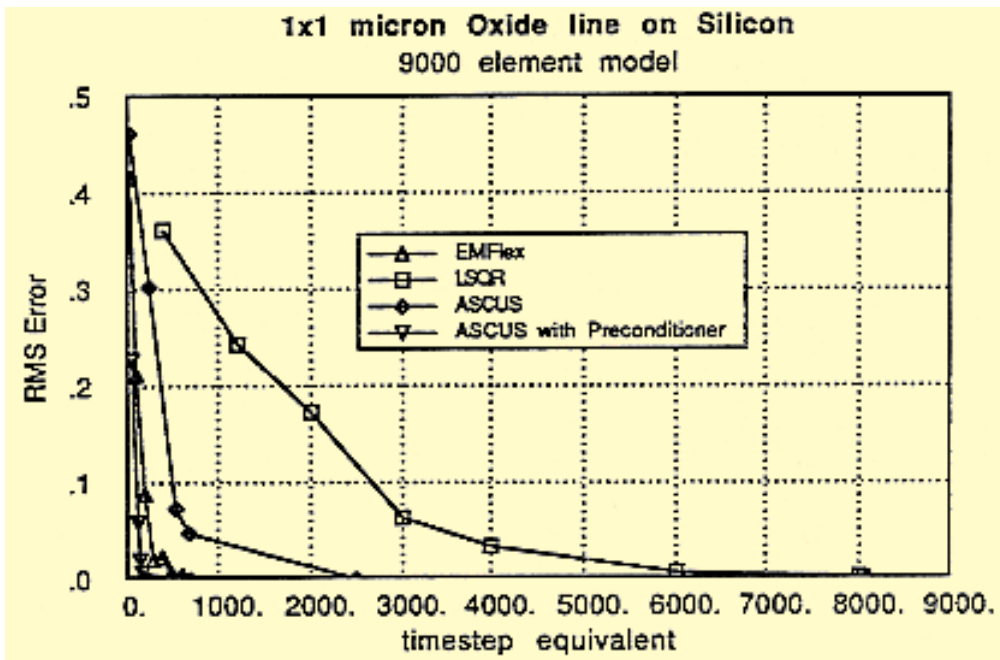


Figure 12. Crossplots illustrating the relative convergence of time- and frequency-domain solvers for the model shown in Figure 8.