<table>
<thead>
<tr>
<th>Title</th>
<th>Overview of Large-Scale Computing: The Past, the Present, and the Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Chew, WC; Jiang, L</td>
</tr>
<tr>
<td>Citation</td>
<td>Proceedings of the IEEE, 2013, v. 101, p. 227-241</td>
</tr>
<tr>
<td>Issued Date</td>
<td>2013</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10722/185861">http://hdl.handle.net/10722/185861</a></td>
</tr>
<tr>
<td>Rights</td>
<td>Creative Commons: Attribution 3.0 Hong Kong License</td>
</tr>
</tbody>
</table>
Overview of Large-Scale Computing: The Past, the Present, and the Future

This paper provides a review of the development of computational electromagnetics, summarizes achievements, and discusses challenging issues and future possibilities.

By Weng Cho Chew, Fellow IEEE, and Li Jun Jiang, Member IEEE

ABSTRACT | This is a brief review of the development of computational electromagnetics (CEM) to partially summarize its achievements, issues, and possibilities.

KEYWORDS | Computational electromagnetics (CEM); finite element method (FEM); integral equation (IE); large-scale computing

I. INTRODUCTION

Calculation is indispensable in many fields. This is even more so in science and technology where a huge number of calculations are routinely done in order to achieve the design goal, gain physical insight, and reveal and predict the inner truth of nature. Hence, instruments to speed up calculations have been in demand since ancient times. First, there is the abacus since historical times, a mechanical device that can retain long numbers, and enables one to perform calculations mechanically. The user of the abacus memorizes the operational moves so that the calculation can be done as “robot-like as possible. In this case, the robot is the trained human [1].

Even with these inventions, calculations were painfully slow. The value of PI was first computed to sixth decimal place by ZU Chong Zhi (祖沖之) in the 5th century [2]. The lower and upper bounds to the value of PI were obtained by using inscribing and excising polygons of a circle with over 12,000 sides. Even then, the calculation took two generations, the work of the father and son team.

Mechanical calculations were in great demand during the industrial revolution (1500s onward) in the United Kingdom. Hence, the log table was invented by Napier, and the slide rule was invented by Oughtred (1600s) [3], [4]. Pascal and Leibniz also invented mechanical calculators in the 1600s. In 1801, the Jacquard loom used the idea of punched cards to control weaving patterns of a loom [5]. Complex weaving patterns could be altered in the loom by swapping punched cards. It allowed complex patterns to be woven by a single machine, and patterned cloth could be produced in large volume. It was the precursor to programmable computing. The 1800s also saw the rapid development of mechanical calculators, the most notable of which was the Babbage machine that could hold seven numbers of 31 decimals each [6], [7].

Calculus was invented in the 17th century by Newton (1642–1727) and Leibniz (1646–1716) [8]. The advent of calculus was followed by the development of many elegant partial differential equations (PDEs) of physics for describing the physics of solid mechanics, elastodynamics, and fluid mechanics. Much of the early works were due to Euler (1707–1783), Lagrange (1736–1813), and Cauchy (1789–1857) [9]. Subsequently, PDEs were extended to describe the physics of electromagnetics (Maxwell, 1831–1879; Heaviside, 1850–1925; Hertz, 1857–1894)
In the 19th century and the early 20th century, we saw the design of ships of unprecedented sizes, and cannons of unmatched power due to improved knowledge of fluid dynamics and solid mechanics [14]. Even Fourier (1769–1830) was reputed for his analysis of heat flow in a cannon [15]. Due to the low velocity of ships, the fluid flow was the simpler Helmholtz (1821–1894) flow or Stokes (1819–1903) flow with less significant dynamic behavior [16]. This period also coincided with the rise of the United Kingdom as a colonial and maritime power [17].

The Second World War saw another push for the demand of computing. To design aircraft and bombs, high-speed projectiles called for even more elaborate solutions of equations of fluid dynamics. These equations are nonlinear, and difficult to solve. However, it was well known that those who could solve these equations would design the best rockets and aircraft, gaining military supremacy. Bevies of women were hired to perform laborious mechanical calculation. They were actually called “computers.”

These unwieldy efforts made it clear that it was necessary to design machines to perform these laborious calculations rapidly [18]. Hence, the first primitive computers were developed around and after the Second World War. The famous ones were the Eniac (1946) and the Illiac (1951) [19], [20].

Before the invention of the computer, engineering and physical insights were obtained by solving geometries with closed-form solutions, such as spheres, cylinders, planes, etc. Often, asymptotic techniques were developed to elucidate the physics from these closed-form solutions. When solutions to more complex geometries were needed, they were obtained by approximate methods, such as the perturbation method and the asymptotic method [21].

Modern technologies have progressed with a breakneck speed. The advancement is possible due to the concurrent synergy of knowledge in materials, process technology, systems, theory, and computation. The availability of the Internet facilitates this synergy at the global level. There is a significant increase in participation of emerging economies in knowledge creation activities in the area of computing. Also, there is a huge market demand for enabling technologies that will change people’s lives. In modern days, theory is almost synonymous with computation. Moreover, the globalized economy has sharply reduced the cost of computing equipment, hence, greatly reducing the cost of computing activities [22].

With the leaps and bounds progress of computer technologies, computing becomes an instrumental component of science and technology development. While gigaflops technology was common place ten years ago, teraflops and petaflops technologies are routinely being touted now [23]. Even at the personal computing level, multicore architecture is now pervasive. The vast popularity of search engines has driven the use of computer clusters or “farms” to meet the speed expectation of users. This has recently evolved into the concept of cloud computing [24]. The rapid growth of visualization technology has given rise to low-cost graphics technologies due to the economy of scale. It has also spurred the interest of using graphic cards or graphical processing units for large-scale computing at a reduced cost [25]. Moreover, the rapid advancement in microelectronics and nanoelectronics has greatly reduced the cost of memory, allowing larger problems to be solved and larger data sets to be stored.

Due to the pervasive use of computers, computing has replaced pencils and papers as the norm in scientific and engineering analyses. Computer visualization has added a new dimension to analyses that are not available in pencil and paper analyses. Moreover, the mushroom growth and proliferation of computer software have made the field more vibrant than ever. Graphical user interface has enabled low-level engineers to perform analyses at a high level, using sophisticated computing tools.

In electromagnetics, while dense matrix systems with millions of unknowns have been solved a decade ago [26]–[28], now over several hundred millions and a billion unknowns can be solved [29]–[31]. These problems are solved with the synergy between hardware improvements, and advances in algorithm design. For sparse matrix systems, and static problems, even larger problems have been solved [29]–[32].

II. DIFFERENTIAL EQUATION SOLVER (DES)

Shortly after the invention of calculus, PDEs were developed to describe the time evolution of physical fields. The space derivative of a field can be taken as the rate of change of the field with respect to a given direction, and likewise for derivatives in other directions and for the time derivative. Luckily for us, the physical behavior of many fields can be understood by studying such derivatives. Hence, differential equations and PDEs were used since the 1700s, and even up to modern days in quantum mechanics.

Such equations were classified as elliptic, parabolic, or hyperbolic [33], [34]. They describe fields of different physical types. An example of a PDE which is elliptic is the Laplace equation or the Poisson equation. They are used pervasively in low-frequency electromagnetic field, and for transport problems in electronics and electrochemistry. They are characterized by the fact that the field or potential associated with such equations has no singularities (away from the source point), or that they are smooth.
there cannot be any singularity propagation in such equations

$$\nabla^2 \phi(r) = -\rho(r)/\varepsilon. \quad (1)$$

Examples of parabolic equations are the Schrodinger equation and the diffusion equation. These equations have the first derivative in time and second derivatives in space. They are characterized by the one-way equation in that the solution only marches forward in time (or backward in time). They can only accommodate one-singularity propagation at a given time. The Schrodinger equation is highly important in quantum mechanics, while the diffusion equation is important for heat transfer and the low-frequency electromagnetic field in conductive media. A typical Schrodinger equation has the form

$$-\frac{\hbar}{2m} \nabla^2 \phi(r) + V(r)\phi(r) = -i\hbar \frac{\partial}{\partial t}\phi(r). \quad (2)$$

A typical diffusion equation is of the form

$$\nabla^2 \phi(r) - \frac{1}{e\tau} \frac{\partial}{\partial t}\phi(r) = 0. \quad (3)$$

An example of a hyperbolic equation is the wave equation. It has second derivatives both in space and time. It can accommodate simultaneously the forward propagating and backward propagating waves, or a wave in forward time and reversed time. It can have two singularities propagating at a given time

$$\nabla^2 \phi(r) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2}\phi(r) = 0. \quad (4)$$

Differential equations are solved commonly by two methods: the finite difference method [35], [36], or a subspace projection method such as the finite element method (FEM) [37]–[39], and the pseudospectral method [40]–[42]. In the finite difference method, a set of grid points is defined, and the derivatives are approximated over such grid points. A matrix system that can be solved easily is thereby obtained.

For the subspace projection method, a set of basis functions is defined to approximate the given field [44]. Since the set of basis functions is finite, it spans a subspace of the larger space that the field is defined over. In such a manner, the PDE can be easily converted to an ordinary differential equation (ODE) in time only. The time derivatives can further be approximated by finite difference, or the subspace projection method, so that the equation can be solved by time stepping or marching. Alternatively, one can Fourier transform in time to remove the time derivatives to obtain a matrix equation that can be solved by an iterative or inversion

$$Lf = g, \quad f = \sum_{n=1}^{N} a_n b_n = b \cdot a$$

$$Lb \cdot a = g, \quad (t, Lb) \cdot a = (t, g)$$

$$\mathbf{L} \cdot a = g, \quad \mathbf{L} = (t, Lb), \quad g = (t, g). \quad (5)$$

Since derivatives only draw information from two points in space infinitesimally spaced apart, the approximations of the derivatives on the computer also draw upon two points close to each other. Hence, only near-neighbor information is needed in forming or approximating derivatives. Hence, the matrix system associated with a differential operator is sparse. If space is approximated by \( N \) space points, the number of matrix elements is proportional to \( N \), as only near-neighbor information is used. Hence, differential equations can be solved easily using the sparse matrix method. A matrix–vector product (MVP) will entail \( O(N) \) operations. Therefore, the matrix system in the differential equation solver (DES) can be solved efficiently by an iterative method. When direct methods are needed, they can be solved by the nested dissection ordering method or the frontal method [43]. When time evolution of the field is needed, one can use the time-marching method.

One would like to solve the problem as expeditiously as possible. However, for wavelike problems, a physical condition such as the Courant stability condition forbids the use of large time steps [45]. The MVP of the differential operator conveys information only to its nearest neighbor grid point, since it is entirely a local operator. Hence, in the time-stepping method, it takes at least the time to propagate the information across the entire simulation region before the quiescent solution is reached. By the Courant condition, the time step and the space step are linearly proportional to each other and are related by the wave velocity. Hence, in the 3-D space, if the number of grid point is \( N \), the length of the simulation region is roughly proportional to \( N^{1/3} \). Hence, it takes \( N^{1/3} \) to propagate information across the simulation region. Therefore, it takes at least \( N^{1/3} \) time to solve the problem. When the structure is resonant, the wave will bounce around multiple times before the quiescent is reached. In this case, the number of iterations can be large. The computational complexity is then \( MN^{1/3} \), where \( M \) is the number of times that the wave will bounce around in the structure. It is proportional to the \( Q \) of the resonant structure. Consequently, the numerical solution of a high-\( Q \) resonant structure is a challenging problem. When iterative solvers are used to solve the sparse matrix system, the same physical
condition applies, and the number of iterations is proportional to $N^{1/3}$, and the computational complexity of solving the matrix system for a wavelike problem is proportional to $N^3$.

For a Laplace or Poisson system, where the velocity of light is assumed infinite, such a causality relation is not necessary. Hence, the information can be sent across the simulation region instantaneously. In the multigrid method, in order to achieve this, a hierarchy of fine and coarse grids can be constructed where information is passed back and forth between the grids. The coarse grid allows the rapid propagation of information across the region, while the fine grid retains the accuracy of the solution [46].

**III. INTEGRAL EQUATION SOLVER (IES)**

In a linear differential equation system, one can define a point source response called Green’s function. Then, by the principle of linear superposition, the field due to an arbitrary distributed source can be obtained by the spatial convolution of the distributed source with Green’s function. Using this concept, an equivalence principle (see Fig. 1) [47], the field in a given region can be expressed as Green’s operator acting on the sources. Hence, integral equations can be obtained rather than differential equations.

Integral equations have the advantage that the unknowns are supported by surface unknowns only, or by volume unknowns in a finite region in space [21], [44], [50]. Hence, the number of unknowns can be much smaller than those in DESs whose unknowns are fields that pervade all of space. Moreover, in integral equations, by the appropriate choice of Green’s function, the radiation condition is automatically satisfied. This is not the case in DESs where the radiation condition has to be replaced by absorbing boundary conditions, or boundary integral equations.

Integral equations can also be converted to matrix equations by the subspace projection method [44], [48]–[50], where the integral operator can be converted into a matrix operator. Since Green’s operator is not a local operator, the matrix representation of Green’s operator corresponds to a dense matrix system. Hence, the storage and operation such as MVPs with a dense matrix system can be computationally expensive.

Despite the dense matrix system, a slew of methods have been developed to solve the dense matrix equation. This includes fast Fourier transform (FFT)-based methods, fast-multipole-based methods, rank-reduction methods, the nested equivalence principle algorithm, recursive algorithms, and so on [21].

The FFT-based methods are efficient when applied to the volume integral equation where the unknowns are densely packed in a volume. When combined with iterative solvers, it allows an MVP to be effected in $N \log N$ operations. However, when applied to surface scatterers, there could be a lot of zero padding, and efficiency can drop. For instance, for a flat, almost spherical scatterer of an arbitrary shape, the surface unknowns $N_s \propto D^2$ where $D$ is the characteristic diameter of the scatterer. But the FFT unknowns have to fill the volume in which the surface scatterer sits. Hence, the volume FFT unknowns $N_v \propto N_s^{1.5}$. Consequently, if the surface unknown number is $N_s$, the computing time scales as $N_v \log N_v \propto N_s^{1.5} \log N_s^2$. For very flat scatterer that sits in a very thin volume, the number of volume unknowns $N_v \propto N_s$. In this case, the FFT-based algorithm can be very efficient. Hence, FFT methods are good for flat surface scatterers, or densely packed volume scatterers. The fast-multipole-based methods, on the other hand, provide low complexity for surface scatterers, irrespective of the shape of the surface scatterers. For low frequency or Laplacian problems, they can provide $O(N)$ complexity for an MVP. For dynamic problems where wave physics is important, the multilevel fast multipole algorithm (MLFMA) has been developed, and it can provide $O(N \log N)$ complexity for an MVP. MLFMA remains to be the only algorithm that can deliver $O(N \log N)$ complexity for an MVP for wave physics scattering problems [21]. Time-domain version of this algorithm has been reported in [51].

![Fig. 1. Illustration of the equivalence principle: in the top case, the scattered field is due to induced currents on the scatterer. But the induced currents that generate the scattered field can be replaced by equivalent currents on a surface that generate exactly the same scattered field. In the bottom case, the incident field on a scatterer is generated by a source outside the scatterer or at infinity. But the incident field on the scatterer can be generated exactly by equivalent currents on a surface.](image)

**IV. MULTISCALE PROBLEMS**

Multiscale problems present themselves in circuits, packages, and chips at all levels of complexity. They are

---

2It is to be noted that for $\alpha > 0$, $\log N \ll N^\alpha$, $N \to \infty$. Hence, $\log N$ grows slowly with $N$.  

---

also found in antennas on complex platforms as well as in nano-optics and nanolithography. While the functioning of the antennas relies on the intricate physics of the feeds, the probes, and the waveguide and resonant structures, the platforms they are mounted on are a lot simpler. Hence, solutions to multiscale problems are instrumental for a large number of applications [55]–[61].

Researchers with various focuses have very different definitions for low frequencies and high frequencies. Fig. 2 shows the one used in this paper. Based on the electrical size, the frequency band is divided into low, mid, and high frequencies. However, it is more universal to use the dominant physics to characterize the problem. These are circuit physics, wave physics, and ray physics from the classic theory point of view, as shown in Fig. 2. When the object’s feature dimension is a small fraction of the wavelength, circuit physics is dominant. The field is smooth and evanescent. Static or quasi-static approximations can be applied to build the model and improve the efficiency. Traditional integrated circuit (IC) physical layer parasitic extraction and PCB modeling are primarily in this domain. When the feature size is over many wavelengths, ray physics is dominant. High-frequency asymptotic approximation can be conveniently applied in the area. In between, the field is oscillatory and propagating. The wave physics has to be understood based on the first principles (Maxwell’s equations) rigorously.

It is critical to evaluate the multiscale structures relative to the wavelength to determine if circuit physics, wave physics, or optics physics is critical to the problem solving. It can effectively help to avoid or identify ill-conditioned numerical systems (Fig. 2). When multiscale structures exist, the meshes that describe their geometries are disparate in sizes, giving rise to very ill-conditioned matrix systems. When mesh sizes are too small compared to wavelength, low-frequency breakdown of the solutions also occurs. Hence, special treatments have to be sought to overcome these problems.

When the frequency is low, electromagnetic physics becomes circuit physics where there are inductances, capacitances, and resistances, which are decoupled from each other. In an electric field integral equation (EFIE), the electric field dominates over the magnetic field. Hence, at low frequency, the inductance physics is swamped by the capacitance physics, giving rise to low-frequency breakdown. This problem can be overcome by the loop-charge (tree/star) decomposition [62]–[68]. In this method, the current on an object is decomposed into loop currents which have zero divergence, and charge currents which have nonzero divergence. The charge currents are either described by tree currents or star currents. When the current is partitioned into these currents, the capacitance physics and the inductance physics of the problem can be separated. But once the physics can be partitioned into different matrix blocks, frequency normalization can be used to rectify this problem.

The disadvantage of the loop-charge decomposition is the need to search for the loops which can be quite unwieldy for highly complex structures. However, it has been known that the circuits community never had to search for loops in solving low-frequency circuits problems. The augmented EFIE (A-EFIE) [60] is motivated by such an observation. In A-EFIE, the EFIE is converted into one that looks like Kirchhoff voltage law, and Kirchhoff current law by augmenting the EFIE with the current continuity equation. In this manner, plus proper frequency normalization, the equation is converted into a saddle point problem that is often used in circuits. The A-EFIE has been found to be low-frequency stable down to statics.

Multiscale problems are also fraught with mixed physics problems. When geometry sizes are small compared to wavelengths, circuit physics predominates, but when geometry sizes are on the order or larger than wavelengths, wave physics rules. Hence, in a multiscale geometry, coexistence of circuit physics and wave physics makes their simulation even more difficult. In such geometry, close interactions are dominated by circuit physics, while far interactions are dominated by wave physics.

To overcome this problem, the equivalence principle algorithm (EPA) [57], [58] was proposed. This algorithm evolved from the nested equivalence principle algorithm (NEPAL) or the recursive Green’s function method [69], [70] for reducing the computational complexity of solving complicated scattering problems. However, the key intent of EPA is to separate wave physics from circuit physics in multiscale problems. By so doing, ill-convergence of matrix equations can be circumvented.

V. WAVELETS SCHEME

The wavelets scheme has been proposed to sparsify dense matrix systems that follow from integral equations [71]–[78]. This is especially effective for circuit physics problems where Laplacian and static fields predominate.

![Fig. 2. The multiscale structures relative to the wavelength. At different scales, different dominant physics will be involved.](image-url)
The nature of the static field is that it does not carry information over long distances [79]. This is confirmed by the fact that the higher the order of the multipole, the faster the algebraic decay of the field. Hence, if one is at a certain distance from the source, the most dominant field is the monopole field, followed by the dipole field, and the higher order multipole fields. Therefore, there is little information in the field far away from a static source. The off-diagonal blocks of the matrix representation of Green’s operator represent the interaction between well-separated groups of sources, and by the above reasoning, they are low rank. Wavelets basis can be used to compress the matrix system that results from integral equations at low frequencies, because the off-diagonal blocks for the matrix system do not contain much information, and hence are low rank. It has been shown that matrix systems from static field can be compressed from $O(N^2)$ to $O(N)$ or $O(N \log N)$ elements using wavelets transforms.

However, for wave physics problems, the wave conveys information over long range. Hence, the off-diagonal blocks contain information for long-range interactions regardless of how well separated the sources are. Even though the long-range interaction can be compressed by wavelets transforms, the long-range interactions cannot completely be compressed. Hence, in the wavelet transformed space, there is always a block of dense matrix representing the $O(N^2)$ long-range interaction that persists in the system regardless of the basis used. Hence, wavelets scheme cannot be used to reduce the computational complexity of wave physics problems, but it can be used to reduce the computational complexity of circuit physics problems.

Moreover, on a smooth surface, a wave problem will induce oscillatory currents, where at least two points per wavelength are needed to capture the embedded information. Therefore, the current cannot be “smoother” than the “Nyquist barrier” of two points per wavelength. But for a long-wavelength problem, the current induced on a smooth surface is infinitely smooth, requiring few sampling points to capture its information content.

VI. WELL-CONDITIONED SCHEMES—CALDERON PROJECTION METHOD

Because of the ill-conditioned nature of many integral equations, the Calderon projection method has been proposed to derive well-conditioned equations from ill-conditioned ones [80]–[82]. The integral operator corresponding to EFIE is unbounded and ill conditioned when the frequency is low. It is also the first kind of an integral operator. However, the square of the EFIE operator can be written as an identity operator plus a compact part. Hence, the EFIE operator is its own preconditioner. This concept has been exploited by a number of researchers to derive well-conditioned systems. While on paper the EFIE operator is its own preconditioner, this fact does not materialize in the approximate numerical representations of the operator. The use of analytic cancellation has been used to achieve this preconditioner property. Recently, the use of curl conforming basis functions has facilitated the numerical implementation of this concept, and numerous papers have been published on this topic [80]. The method still suffers from low-frequency inaccuracy problems, and the loop-charge decomposition concept has been used to alleviate this problem. One of the remaining issues regarding this method is the inversion of the Gram matrix for complex loop-charge systems, which can be quite complicated. Recently, the perturbation method has been used to alleviate the low-frequency inaccuracy problem in this method [83].

VII. DIVIDE-AND-CONQUER SCHEMES

Even though computers have advanced dramatically in the computing speed and accessible memory, the practical complexities of real electromagnetic problems still make $O(N^2)$ and $O(N^3)$ intolerable. As a general algorithm development paradigm, divide and conquer (DAC) is an important process to reduce the overall computational cost down to a reasonable level. In general, DAC recursively divides one big problem into smaller subproblems. By collecting contributions from all subproblems in a certain order, the final solution can be obtained with a much lower computational cost. Because the computational load is distributed into subproblems, a distributed hardware system could be naturally employed to support the DAC strategy. Hence, many algorithms developed on supercomputers or parallelized clusters use DAC ideas.

One way to implement DAC algorithms is to partition the objects based on their spatial distances. Another common approach is to partition the resultant matrix equation based on the rank deficiency. Both of them rely on the decay (at various rates) of electromagnetic (EM) waves versus the distance.

One of the early DAC methods was the Barnes–Hut algorithm in 1986 [84]. It was for calculating $N$ particle interactions via the gravitational force field. Because it calculates the interactions of each particle with all well-separated groups, its complexity is in $O(N \log N)$. Appel’s algorithm [85], [86], which was published one year earlier than the Barnes–Hut algorithm, has actually achieved $O(N)$ complexity because it carefully considered interactions between well-separated groups instead of interactions between particles and groups. Both of them used the center-of-mass (CoM) concept. It can be conveniently extended to the center-of-charge (CoC) idea in the static field [87]. Hence, both methods have been employed to solve parasitic capacitance extractions [87], [88]. It was also proven that the CoC method is equivalent to the second-order fast multipole algorithm if the CoC overlaps with the origin of the coordinate system [87].

The fast multipole algorithm (FMA) was one of the great breakthroughs of computational electromagnetics.
is greatly decreased. (a) Direct interaction computation among objects requires \( O(N^2) \) operations. (b) Hierarchically organized interactions by groups reduce the overall computation cost greatly to \( O(N \log N) \) or less.

Because in the midfrequency propagating wave dominates, plane waves instead of multipoles are used to implement the full aggregation, translation, and disaggregation scheme. Different from static FMA and LF–FMA who have the dense translators, MLFMA has diagonal translators. Because the plane wave sampling rate proportionally increases with bigger box sizes (which does not happen for low frequencies), interpolation and anterpolation were developed to optimize the computational cost and storage requirement. They are the key technologies that successfully upgraded FMA to MLFMA and guaranteed the \( O(N \log N) \) computational cost per MVP for wave physics problems [21], [103].

When the frequency drops, evanescent waves instead of propagating waves will be dominant. Hence, MLFMA using propagating plane waves has the low-frequency breakdown problem when the translation distance is below 0.1 wavelengths [54], [104]. To solve the low-frequency breakdown of MLFMA and the high-frequency inefficiency of LF–FMA, the mixed-form FMA (MF–FMA) was developed to provide a unified broadband FMA solution [105], [106]. It employs an analytical transformation between low-frequency multipoles and midfrequency plane waves to adopt both low- and high-frequency physics in the same FMA oct-tree. Hence, it can handle broadband MVP without frequency-dependent breakdowns.

One drawback of the FMA method is its kernel dependence because it relies on the multipole expansion or the plane wave representation of Green’s function. A kernel-independent FMA developed by Ying [107] uses the continuous equivalence source on the surface of the enclosing box to replace the analytic multipole expansion. The translations from a source box to its interaction list boxes are done using FFT. This method is very similar to precorrected FFT (PFFT) [89], AIM [93], and the wavelet method by Wagner and Chew [76].

Many successful DAC methods started with the static field analysis. The PFFT [89] was first proposed for FASTCAP, the well-known capacitance extraction code developed by White’s group at the Massachusetts Institute of Technology (MIT). The differences between PFFT and FMA are shown in Fig. 4. PFFT partitions the charge distribution by uniform boxes. Then, the charges in each unit box are replaced by equivalent charges located at the uniform grids of the cell box surface. Then, the potential evaluation of the uniform grid charges is a convolution integral which can be efficiently computed by the FFT method. Because equivalence sources are not accurate for nearby interactions, direct computations are required to correct the near-field data. The overall computational complexity is \( O(N \log N) \). PFFT was further applied to various full-wave problems [90]–[92]. Almost in the same period of time, the AIM, which is very similar to PFFT, was directly developed for the full-wave scattering problems [93]. AIM separates the field evaluation into far field and near field from the very beginning. Then, the auxiliary...
basis functions are used to decompose the source and evaluation points onto uniform grids. The resultant far-field interaction matrix becomes Toeplitz that can be solved by FFT. The near-field interaction is obtained by a direct calculation. AIM has been successfully applied to large-scale scattering problems of several wavelengths. Beyond metallic surface scattering problems, it has been extended to volume integral equation related methods to solve composite dielectric and conducting objects [94], arbitrary shaped magneto–dielectric objects [95], and parametric geometries [96].

Another type of DAC methods employs the rank deficiency of the matrix system obtained from the EM modeling. In the IES3, the matrix representation of the integral equation operator was first obtained using the method of moments (MoM) [97]. Then, the off-diagonal submatrices are assumed to be of lower ranks than their dimensions. Using a precomputed rank map, the matrix is divided into many submatrices whose ranks are further compressed by eliminating unimportant modes through the singular value decomposition (SVD). Hence, the total computational cost needed for the MVP is greatly reduced. Later, a similar method named the predetermined interaction list oct-tree (PILOT) QR algorithm was developed [98]. Instead of using the precomputed rank map as in IES3, PILOT employs the fast multiple oct-tree to organize the interaction list. The setup time is greatly reduced. Both methods have been applied to low-frequency modeling situations with the complexity of $O(N \log N)$ or $O(N \log^2 N)$. One significant advantage of these techniques is that they are kernel independent. The rank deficiency was also successfully implemented in a multilevel matrix decomposition algorithm (MLMDA) [99]. It uses a reduced set of equivalence sources and observation points to achieve the rank deficiency. Then, it develops a well-designed butterfly-like submatrix–vector multiplication process. Compared to AIM $O(N^{1.5} \log N)$ for the 2-D surface scatterers, MLMDA does not need the FFT on the entire grid, and its computational complexity is $O(N \log^2 N)$.

The hierarchical matrices (H-matrices) method is a method recently applied to CEM applications [108]–[111]. It assumes that typical integration kernels are asymptotically smooth. Then, the panel clustering method is applied to obtain the kernels’ degenerate approximations. If all admissible sub-blocks (well separated by definition) of the matrix have the rank less than certain $k$, the matrix is defined as the H-matrix structure. Then, all matrix–vector multiplications, additions, and inversions can all be handled in lower ranks hierarchically. It is kernel independent and has been applied to solve large-scale IC parasitic problems. In principle, this method is valid for the low-frequency regime where the quasi-static approximation is valid because its efficiency comes from a systematic rank deficiency exploitation strategy.

The adaptive cross approximation (ACA) method employs similar rank deficiency concept to construct an approximated matrix without computing the whole matrix [112]. Unlike the H-matrix method that uses submatrices, in its early stage, ACA was proposed to approximate the whole matrix. Later it was developed into a hierarchical multilevel algorithm [113]. It groups unknowns hierarchically based on their geometrical relations. Then, the impedance matrix is split into coupling sub-blocks. Most off-diagonal submatrices corresponding to well-separated interactions will be compressed by the original ACA method. The diagonal matrices and neighboring interactions will be computed directly. This method has been successfully applied to static and dynamic problems.

Most of aforementioned algorithms are used to accelerate the MVP in the numerical calculation. However, there is another category of DAC methods that influences the initial physical model setups. It is the domain decomposition method (DDM) that divides the problem into regions, normally larger than those of DAC methods we already discussed. The mathematical formulation is directly based on this partition. It does not group basis to reduce the computational cost at this stage even though all aforementioned methods could be applied to accelerate its computation.

One method that has been developed is the equivalence principle algorithm (EPA) [57], [58]. It is related to the earlier method of nested equivalence principle algorithm (NEPAL) and the recursive Green’s function method [69], [70]. To implement EPA, the objects are partitioned by artificial equivalence surfaces (Fig. 5). Using the tangential electric field and the magnetic field, equivalent electric current and magnetic current are created. Based on Huygens’ principle, interactions between objects are replaced by internal interactions inside each equivalence surface and intermediate translations between equivalence surfaces [57]. It is similar to the [S] parameter with ports placed all over the equivalence surface. The EPA method has been applied to midfrequency problems such as
antenna arrays and large-scale scattering problems [58] (Fig. 6). Its low-frequency extension is being studied [114], [115]. This method can be regarded as the domain decomposition for IES, and NEPAL [69] is related to the nested dissection algorithm [116] for DES.

Yet another successful DDM development is the non-overlapping DDM developed by Lee’s group [117]. It directly segments object bodies during the partition process. Its difference compared with the EPA method [57] is illustrated in Fig. 7. Then, the first-order or second-order Robin-type transmission conditions are established to enforce the continuity of boundary conditions at the cutting surface. Later, a cement technique is employed to adapt to nonoverlapping meshes. It has been successfully applied to analyze both large-scale scattering and coupling problems [117]–[119]. Very impressive results have been achievement from these methods. One example is shown in Fig. 8.

Another DAC scheme to note is the diakoptic approach. It has also been generalized to using the equivalence principle and FEM [120]–[131].

For multilayer systems, DDM can be naturally applied by splitting regions by layer interfaces. The time-domain layered finite element reduction recovery (LAFT–RR) method in [132] rigorously reduces the matrix of the multilayer system into that of a single-layer system irrespective of the original problem size. It is able to solve a single-layer matrix problem in $O(M)$, where $M$ is the number of single-layer unknowns. It is very suitable for an on-chip structure analysis.

Recently, another very popular time-domain DDM is the discontinuous Galerkin time-domain method (DGTD) [133], [134]. It can handle various elements with different shapes, accommodate nonconforming meshes, achieve high-order accuracy, and favor parallelization process [135]–[137]. A detailed review paper on this method is given by Chen and Liu in this special issue [138].

VIII. LAYERED MEDIUM FMA

While it is relatively easy to develop fast methods for free space and the homogeneous medium, it is more difficult to
develop them for the layered medium [52]–[54]. When the layered medium is thin, a thin stratified medium fast multipole algorithm (TS–FMA) has been developed [52]. In this method, the layered medium is assumed thin, and the observation point is close to the horizontal. In this case, the Sommerfeld integration path for the layered medium Green’s function can be detoured to the vertical branch cut, and be expressed as an efficient integral summation of 2-D Green’s functions. Then, 2-D MLFMA can be applied to accelerate the computation of MVPs.

When the layered medium is not thin, the fast inhomogeneous plane wave algorithm (FIPWA) can be applied [53], [54]. The advantage of FIPWA is that it keeps the important form of Sommerfeld integral intact and yet allows the building of the factorization of Green’s function into the complex layered medium Green’s function.

There have been attempts of using discrete complex image methods (DCIM) to accelerate the calculation of layered medium Green’s functions [139]–[141]. DCIM uses the complex images to obtain the approximated closed forms of integrals with the layer medium Green’s function. But it has difficulties in processing the poles and branch-point singularities. Also it has unpredictable errors for the far-field interaction. Recently, a Sommerfeld branch cut (SBC) method was proposed to capture these singularities [142]. DCIM was also combined with the static layer medium fast multipole algorithm to accelerate the setup time for translators [143].

**IX. HIGH-PERFORMANCE COMPUTING (HPC) FOR CEM**

Computer hardware was advancing when CEM was progressing. High-performance computing for CEM has become an unavoidable research direction adopted by the CEM community. Using distributed central processing unit (CPU) capability and distributed memories, massive computational workload can be shared. Hence, significant speedups can be achieved by the hardware scaling.

There are three major types of HPC platforms for today’s CEM researchers.

1) Computer clusters: it is very economical to build a cluster with 10 or 20 nodes to achieve up to 10–20 times acceleration of the original computation task.

2) Supercomputers: such as those in the top 500 supercomputer list [144]. They are highly optimized computing platforms with tens of thousands of nodes.

3) Cloud computing: it is a technology that is still in its infancy today. It emphasizes services and resource sharing instead of scientific computing. Hence, its impact on CEM is still unknown [145], [146].

Parallel computer memory architectures are of three types [147]: 1) shared memory system; 2) distributed memory system; and 3) hybrid distributed-shared memory system that is a combination of the previous two types (Fig. 9). Most today’s supercomputers are the hybrid system.

---

Fig. 8. Microwave photonic crystal (MPC) was simulated using the DDM proposed in [120]. The top figure shows the geometry and domain partition of MPC. The bottom figure shows the negative reflection and the resultant electric field at 9.7 GHz. One thousand forty eight subdomains were required (courtesy of Prof. J. F. Lee).

Fig. 9. A hybrid distributed-shared memory computer memory architecture. Each unit node is a shared memory system—there are two or more symmetric multiprocessors (SMPs) sharing the same memory resources. Then, a distributive system is formed by connecting shared memory nodes through networks [148].
One key challenge of parallelizing CEM algorithms is how to maintain its scalability when the number of nodes is increasing. If there is little communication between computing nodes, the algorithm is said to be embarrassingly (or pleasingly) parallel. Or else, hard work is needed to balance the workload in the parallel system, synchronize processes on all the nodes, and expedite and reduce the internode communications to optimize the scalability.

Large-scale CEM was primarily spurred by the need for radar cross-section (RCS) analysis. Moreover, fast multipole algorithms were very successful in handling that challenge on the single node computer. Consequently, many HPC CEM efforts were focused on the parallelization of FMA related algorithms, which is also academically challenging.

One of the initial efforts to parallelize FMA is the Fast Illinois Solver Code (FISC) at the University of Illinois at Urbana-Champaign (UIUC). Using shared memory system, the RCS of a full-size aircraft at X-band [27] was computed for the first time. Later, it also successfully calculated the solution to a MoM matrix (a dense matrix) with ten million unknowns [27]. A distributed memory parallel algorithm, ScaleME, was later developed at UIUC [28]. For the first time, it achieved the record of ten million unknowns for the RCS calculation. It used 126 processors of the SGI Origin 2000 system to support that calculation. Therefore, many advanced FMA parallelization efforts have been made. Today, for static particle interactions, the record is set to solving three trillion particles in 11 min using FMA [32]. For electromagnetics (dynamic) problems, over half a billion to over a billion unknowns have been possible [29]–[31].

It is very difficult to partition and balance the hierarchical FMA oct-tree among a large number of computing nodes. Improper arrangement easily causes significant data traffic between nodes for FMA aggregation, disaggregation, and translation. Normally, the overall scalability is dramatically degraded when the number of nodes is greater than 32. To overcome it, a technique based on an appropriate partitioning scheme for each level of the MLFMA tree was developed [29]. It partitions both subdomains and their samples using the load-balancing algorithms. As a result, much higher parallelization efficiency was achieved. Up to 374 million unknown RCS problems have been solved based on this method by 2011 [30].

In 2009, another method named FMA–FFT showed extraordinary scalability [31]. It employed the idea of using a simpler solver to achieve higher scalability. It only applied one level FMA to organize irregularly distributed sources. Hence, almost no internode communication is needed for the FMA tree, which was the trouble source of the bad load balancing. Then, the FFT method was used to calculate coarse level interactions. By this, the workload could be evenly distributed to all nodes, and very limited internode communication was needed. Its award winning scalability was almost 100% on 512 nodes [149] (Fig. 10). This method was further extended to the MLFMA–FFT method that employed the multilevel FMA on each computing node and FMA–FFT over the whole distributive system [31].

The FDTD method is one representative embarrassingly parallel algorithm [149], but it needs a larger number of unknowns than integral equation solvers. It has been parallelized by so many researchers in commercial software tools that it is totally beyond authors’ capability to summarize them.

X. CONCLUSION AND FUTURE
CEM has been dramatically advanced by a successful integration of physics, mathematics, and computer technologies. It has also advanced science and engineering knowledge and contributed to the changes of today’s technology landscape. As a necessary tool for electromagnetic analysis and its broad applications, CEM and large-scale computing will continue to evolve. It will enhance our deeper understanding of the physics of highly complex systems.
algorithm from j-iich-calculates-over-
three-trillion-particles.html


ABOUT THE AUTHORS

Weng Cho Chew (Fellow, IEEE) received the B.S. degree in 1976, the M.S. and Engineer’s degrees in 1978, and the Ph.D. degree in 1980, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge.

He served as the Dean of Engineering at The University of Hong Kong, Hong Kong, from 2008 to 2011. He was a Professor and the Director of the Center for Computational Electromagnetics and the Electromagnetics Laboratory, the University of Illinois at Urbana-Champaign, Urbana. Before joining the University of Illinois at Urbana-Champaign, he was the Department Manager and a Program Leader at Schlumberger-Doll Research, Cambridge, MA. He is the originator of several fast algorithms for solving electromagnetic scattering and inverse problems. He has led a research group that has developed parallel codes to solve dense matrix systems with tens of millions of unknowns for the first time for integral equations of scattering. He has authored a book entitled Waves and Fields in Inhomogeneous Media (New York: IEEE Press, 1999), coauthored two books entitled Fast and Efficient Methods in Computational Electromagnetics (Boston, MA: Artech House, 2001) and Integral Equation Methods for Electromagnetic and Elastic Waves (San Rafael, CA: Morgan & Claypool, 2007, 1st ed.), and authored or coauthored more than 300 journal publications, more than 400 conference publications, and more than ten book chapters. His research interests include the areas of waves in inhomogeneous media for various sensing applications, integrated circuits, microstrip antenna applications, and fast algorithms for solving wave scattering and radiation problems.

Chew and Jiang: Overview of Large Scale Computing: The Past, the Present, and the Future
Dr. Chew served on the IEEE Adcom for Antennas and Propagation Society as well as on the Geoscience and Remote Sensing Society. He has been active with various journals and societies. He is the Fellow of the Optical Society of America, the Institute of Physics, the Electromagnetics Academy, the Hong Kong Institute of Engineers, and was the National Science Foundation Presidential Young Investigator. He received the Schelkunoff Best Paper Award for the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, the IEEE Graduate Teaching Award, the UIUC Campus Wide Teaching Award, and the IBM Faculty Awards. He was a Founder Professor of the College of Engineering, and previously, the First Y. T. Lo Endowed Chair Professor in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. From 2005 to 2007, he served as an IEEE Distinguished Lecturer. He served as the Cheng Tsang Man Visiting Professor at Nanyang Technological University, Singapore, in 2006. In 2002, the Institute for Science Information Citation elected him to the category of most highly cited authors (top 0.5%). In 2008, he was elected by the IEEE Antennas and Propagation Society to receive the Chen-To Tai Distinguished Educator Award. He is currently the Editor-in-Chief of the Journal of Electromagnetic Waves and Applications/Progress in Electromagnetic Research Journal, and on the Board of Directors of Applied Science Technology Research Institute, Hong Kong.

Li Jun Jiang (Member, IEEE) received the B.S. degree in electrical engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1993, the M.S. degree in electrical engineering from the Tsinghua University, Beijing, China, in 1996, and the Ph.D degree in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana, in 2004.

From 1996 to 1999, he was an Application Engineer with the Hewlett-Packard Company. Since 2004, he has been the Postdoctoral Researcher, the Research Staff Member, and the Senior Engineer at IBM T.J. Watson Research Center, Yorktown Heights, NY. Since the end of 2009, he has been an Associate Professor with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong. His research interests focus on electromagnetics, EMC/EMI, antennas, multidisciplinary EDA solutions, radio-frequency (RF) and microwave technologies, high-performance computing (HPC), etc.

Dr. Jiang received the HP STAR Award in 1998. In 2003, he received the IEEE Microwave Theory and Techniques Graduate Fellowship Award, and in 2004, the Y.T. Lo Outstanding Research Award. In 2008, he received the IBM Research Technical Achievement Award. He was the Semiconductor Research Cooperation (SRC) Industrial Liaison for several academic projects. Since 2009, he has been the SRC Packaging High Frequency Topic TT Chair. He has been serving as the Technical Committee Member for the IEEE Electrical Design of Advanced Packaging and Systems Symposium (EDAPS) since 2010, the Scientific Committee Member of the 2010 IEEE Workshop on Simulation and Modeling of Emerging Electronics (SMEE), and the Guest Associate Editor of the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION since 2011. He also serves as the reviewer of many primary electromagnetics journals and special session organizer of many international conferences.