A Recompressed Nested Cross Approximation for Electrically Large Bodies

Nathan M. Parzuchowski, *Member, IEEE*, Brenton Hall, Isroel M. Mandel, *Member, IEEE*, Ian Holloway, and Eli Lansey, *Member, IEEE*

***Abstract*— A recompressed nested cross approximation (rNCA) based closely on the recent fast nested cross approximation (fNCA) algorithm is formulated in this article. The proposed method builds on previous work in which the fNCA was formulated in a purely algebraic and kernel-independent fashion, using a top-down recursive application of the adaptive cross-approximation (ACA). Our proposed method employs ACA recompression to avoid the need to compute low-rank approximations of excessively large far-field matrices, and thus mitigates the effects of high-frequency rank growth on run-time scaling for electrically large models. The low run-time and memory cost allows for efficient parallel computation of $\mathcal{H}^2$-matrices for systems of excessive electrical sizes. Radar cross sections (RCSs) are evaluated for electrically large instances of a perfectly conducting sphere and the NASA Almond. We observe near-linear scaling of memory cost and construction time.**

***Index Terms*— Boundary integral equations, linear algebra, radar cross section (RCS).**

I. Introduction

OVER the past few decades, a myriad of acceleration techniques for electromagnetic integral equations solvers have been developed based on the $\mathcal{H}$-matrix data-sparse storage format [1], [2], [3]. Of these, some of the most compelling are algebraic methods which populate the $\mathcal{H}$-matrix directly from the entries of the dense impedance matrix, such as the adaptive cross approximation (ACA) [4], [5]. These methods offer the ability to construct low-rank approximations derived from algebraic decompositions of the full integral operator matrices.

A major challenge for both algebraic and approximate-kernel techniques (see [6], [7], [8]) is the known tendency for the reduced rank of admissible submatrices to grow proportionally with electrical size [9], [10], leading to impractical asymptotic complexities. In the context of the fast-multipole method (FMM), this problem has been circumvented by transforming to a representation where the translation operator is diagonalized [11], [12]. However, this format relies on the fast matrix-multiply to retain a low asymptotic complexity and thus precludes the use of direct solve methods via lower–upper (LU) decomposition.

To address this challenge for algebraic methods, more sophisticated hierarchical formats have been employed, such as the $\mathcal{H}^2$- and directional $\mathcal{H}^2$-matrix [3], [9], [13], [14], [15]. The directional variant of $\mathcal{H}^2$-matrices attempts to rigorously build angular partitioning into the storage format, forcing all admissible sub-blocks to have a reduced rank bounded independently of electrical size. Unfortunately, it is not clear how to formulate an LU decomposition in the directional $\mathcal{H}^2$-matrix format, as the employed directional cluster bases are not closed under multiplication operations. In contrast, the $\mathcal{H}^2$-matrix method does not rigorously resolve the high-frequency rank growth problem, but rather mitigates it by allowing for deeper hierarchies and thus finer grained partitioning. Smaller partitions ultimately mean a more confined angular breadth of field clusters with respect to source clusters, and under certain conditions it can be shown that this puts an upper bound on the reduced rank [10], [16]. In this work, we describe a new method to populate a conventional $\mathcal{H}^2$-matrix directly from the entries of a method-of-moments (MoM) matrix for electromagnetic integral equations.

$\mathcal{H}^2$-matrices present challenges for algebraic fill methods, as they employ global nested bases to encode data sparsity. Naively, direct computation of these bases would entail constructing low-rank approximations of the coupling between each source cluster and its entire far-field; such an approach would be computationally intractable. Instead, a common methodology used to compute these bases, called the nested cross approximation (NCA), uses ACA to generate low-rank approximations of the far-field coupling based on representative sets of mesh elements [17], [18], [19]. How these representative elements are chosen is an open question. In recent efforts [20], ACA was used recursively in top-down fashion, starting with a full computation of the far-field coupling matrix at the highest level, and working downward reusing ACA pivots from parent levels to approximate their contributions to lower levels. This method shows great promise, but ultimately retains the difficulty that the far-field coupling matrices will eventually become intractably large as electrical sizes increase. To resolve this problem, the authors proposed a bottom-up recursive pass to select representative elements for all levels of

Manuscript received 26 April 2022; revised 23 November 2022; accepted 3 December 2022. Date of publication 23 January 2023; date of current version 6 March 2023. This work was supported by the Independent Research and Development (IRAD) funds from Riverside Research. *(Corresponding author: Nathan M. Parzuchowski.)*

Nathan M. Parzuchowski, Brenton Hall, and Ian Holloway are with Riverside Research, Beavercreek, OH 45431 USA (e-mail: nparzuchowski@riversideresearch.org).

Isroel M. Mandel and Eli Lansey are with Riverside Research, New York, NY 10038 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAP.2023.3237277.

Digital Object Identifier 10.1109/TAP.2023.3237277

the hierarchy. In this work, we modify the top-down algorithm to work on one admissible block at a time, rather than using the entire far-field. This obviates the need for the bottom-up step, but cannot achieve the $\mathcal{O}(Nk)$ run-time complexity offered by the existing method. However, this sacrifice offers greater flexibility in memory utilization, presenting clear opportunities for process- and thread-level parallelism, as well as external-memory processing algorithms to circumvent random access memory limitations. These computational techniques enable the solution of problems of dramatic electrical size. In addition, our approach requires fewer levels of approximation.

II. Formalism

In this section, first we give a brief introduction to the $\mathcal{H}^2$-matrix format, followed by a discussion of the NCA. We then summarize the method proposed in [20], which we will refer to as fast NCA (fNCA). Finally, we will introduce our modified NCA, which we refer to as recompressed NCA (rNCA).

A. $\mathcal{H}^2$-Matrix Format

$\mathcal{H}^2$-matrices are a subset of $\mathcal{H}$-matrices, employing hierarchical subdivision to identify *admissible* submatrices that admit a data-sparse representation. While $\mathcal{H}$-matrices make use of local low-rank approximations of admissible blocks, $\mathcal{H}^2$-matrices utilize a combination of global low-rank row/column cluster bases and low-rank coupling matrices of admissible blocks. This combination results in greater overall compressibility of the original dense matrix.

Consider a geometric mesh which has been subdivided via recursive splitting into a binary cluster tree $\mathcal{T}$. Take for example two clusters of row and column basis functions, labeled t and s, respectively, with sizes N_t and N_s. The matrix block formed by t and s admits a low-rank representation if the two clusters reside at the same level of $\mathcal{T}$,[1] and obey a parabolic admissibility condition [9], [10]

$$\kappa(\max\{\mathtt{diam}(t), \mathtt{diam}(s)\})^2 \leq \eta\, \mathtt{dist}(t,s). \tag{1}$$

Here, $\mathtt{diam}$ denotes the largest distance between two elements of a single cluster, $\mathtt{dist}$ is the shortest distance between two elements in separate clusters, κ is the wavenumber taken from the Helmholtz kernel operator, and η is a tunable scaling parameter. Unlike more conventional admissibility conditions [21], e.g.,

$$\max\{\mathtt{diam}(t), \mathtt{diam}(s)\} \leq \eta\, \mathtt{dist}(t,s). \tag{2}$$

Equation (1) uses frequency information in a way that ensures convergence of low-rank approximations of high-frequency kernels. This condition leads to a significantly finer matrix partition, which is only tractable for $\mathcal{H}^2$-matrices due to their reusable global bases.

[1]In principle, clusters at different levels in the hierarchy can be treated as admissible, however this complicates definitions, formulations, and implementation strategies. For the sake of clarity, we impose the restriction to same-level admissibility.

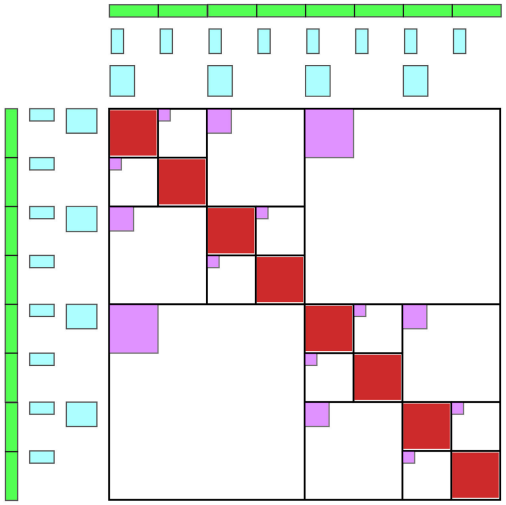


Fig. 1. Schematic $\mathcal{H}^2$-matrix format. Dense blocks are colored red, coupling matrices are maroon, cluster bases green, and transfer matrices light blue (grayscale: listed from darkest to lightest). White space indicates memory reduction. In this example, only diagonal blocks are stored dense, but in practice, some off-diagonal blocks are also dense.

In $\mathcal{H}^2$-matrices, admissible blocks of the MoM matrix are approximated by

$$Z(t,s) \approx V(t) S(t,s) W^T(s). \tag{3}$$

Here $Z(t,s) \in \mathbb{C}^{N_t \times N_s}$ denotes a submatrix of the MoM matrix, $V(t) \in \mathbb{C}^{N_t \times k_t}$ and $W(s) \in \mathbb{C}^{N_s \times k_s}$ are cluster bases for t and s with respective reduced ranks of k_t and k_s, and $S(t,s) \in \mathbb{C}^{k_t \times k_s}$ is the coupling matrix for the t, s block.

To facilitate near linear algorithmic complexities, cluster bases must be *nested*. That is, nonleaf-level cluster bases are represented by matrix products

$$V(t_{1\cup 2}) = \begin{pmatrix} V(t_1)B(t_1) \\ V(t_2)B(t_2) \end{pmatrix} \tag{4}$$

where $t_{1\cup 2}$ is the parent cluster of t_1 and t_2, and $B(t_i) \in \mathbb{C}^{k_{t_i} \times k_{t_{1\cup 2}}}$ are transfer matrices encoding the nesting relationship. A rudimentary schematic example of an $\mathcal{H}^2$-matrix is shown in Fig. 1.

B. Nested Cross Approximation

The NCA gets its name from the employment of ACA to construct nested cluster bases directly from the entries of a matrix. The general approach is to identify a set of k_t proper pivots τ drawn from each cluster t, as well as a set of k_t proper far-field pivots $\bar{\tau}$ drawn from the far-field F_t, where

$$F_t = \bigcup \{ s \in \mathcal{T} | \exists\, t' \supseteq t, s \text{ and } t' \text{ are admissible} \}. \tag{5}$$

The submatrix $Z(t, F_t)$ should be adequately reconstructed by a rank-k_t cross-approximation built with the pivot set $(\tau, \bar{\tau})$. Such a set of proper pivots can be immediately constructed using ACA to compress $Z(t, F_t)$.

Assuming pivot sets have been determined for each cluster in $\mathcal{T}$, we can construct approximations to admissible submatrices via the relationship [17]

$$Z(t,s) \approx Z\big(t,\bar{\tau}\big)\big(Z\big(\tau,\bar{\tau}\big)\big)^{-1} Z(\tau,\sigma)\big(Z\big(\bar{\sigma},\sigma\big)\big)^{-1} Z\big(\bar{\sigma},s\big) \tag{6}$$

where $(\sigma, \bar{\sigma})$ are the proper pivots for the field cluster s. We can extract definitions from this expression

$$V(t) \equiv Z\big(t, \bar{\tau}\big)\big(Z\big(\tau, \bar{\tau}\big)\big)^{-1} \tag{7}$$

$$W(s) \equiv Z\big(s, \bar{\sigma}\big)\big(Z\big(\sigma, \bar{\sigma}\big)\big)^{-1} \tag{8}$$

$$S(t, s) \equiv Z(\tau, \sigma). \tag{9}$$

Error bounds on this approximation were established in [17]. We can see that in the limit that $(\tau, \bar{\tau}) \rightarrow (t, F_t)$ and $(\sigma, \bar{\sigma}) \rightarrow (s, F_s)$, $V(t)$, and $W(s)$ become identity matrices, and $S(t, s) = Z(t, s)$. Nesting relationships are provided by

$$B(t_1) = Z\big(\tau_1, \bar{\tau}_1\big)\big(Z\big(\tau_{1\cup 2}, \bar{\tau}_1\big)\big)^{-1} \tag{10}$$

$$B(t_2) = Z\big(\tau_2, \bar{\tau}_2\big)\big(Z\big(\tau_{1\cup 2}, \bar{\tau}_2\big)\big)^{-1} \tag{11}$$

where $t_{1\cup 2}$ is the parent cluster of t_1 and t_2.

Based on these definitions, it is straightforward to compute the entire $\mathcal{H}^2$-matrix once the proper pivots have been identified. The identification task is challenging, however. As mentioned earlier, direct ACA compression of $Z(t, F(t))$ is prohibitively expensive for large problems, so approximate techniques must be formulated.

C. Fast NCA

We now summarize the method proposed in [20], which applies ACA in top-down fashion to construct nested cluster bases recursively. First, top-level clusters (i.e., the highest level clusters that participate in admissible blocks) are identified. Consider a top-level row-cluster t at depth l. We construct the far-field coupling matrix for the top level

$$A^{(l)}(t) \equiv Z\big(t, F^{(l)}(t)\big) \tag{12}$$

where the same-level far-field is given by

$$F^{(l)}(t) = \bigcup \{s \in \mathcal{T} | s \text{ and } t \text{ are admissible}\}. \tag{13}$$

We now compress $A^{(l)}$ by application of ACA, obtaining $(\tau, \bar{\tau})$ for t. The compressed representation of $A^{(l)}(t)$ can be reused as needed to construct transfer or cluster basis matrices for t.

At the next level, we proceed as before for t's children, t_1 and t_2. However, to build nesting relationships between parent and child cluster bases, we must incorporate far-field coupling information from higher levels. We do so by appending the columns corresponding to $\bar{\tau}$ generated at level l

$$A^{(l+1)}(t_1) \equiv \big(Z\big(t_1, F^{(l+1)}(t_1)\big)\ Z\big(t_1, \bar{\tau}\big)\big) \tag{14}$$

$$A^{(l+1)}(t_2) \equiv \big(Z\big(t_2, F^{(l+1)}(t_2)\big)\ Z\big(t_2, \bar{\tau}\big)\big). \tag{15}$$

Now we may once again apply ACA to extract $(\tau_1, \bar{\tau}_1)$ and $(\tau_2, \bar{\tau}_2)$ from $A^{(l+1)}(t_1)$ and $A^{(l+1)}(t_2)$, respectively. We continue down $\mathcal{T}$ in this fashion until proper pivots have been computed for every cluster. An analogous procedure is used to construct the column cluster bases.

D. Recompressed NCA

To motivate the need for a novel NCA fill method, we note that the computational cost for the fNCA algorithm can be gleaned from the size of $A^{(l)}(t)$, with $N/2^l$ rows and $(N/2^l)C_{sp} + k$ columns in general. Here, k is the cluster basis rank, N is the total number of unknowns, l is the level of the cluster tree, and the sparsity constant C_{sp} is the largest number of admissible blocks any single cluster participates in. Taking into account the $(n + m)k^2$ cost for ACA compression of an $n \times m$ matrix, and summing all levels in the $\mathcal{H}^2$-matrix, we arrive at an expression for the total cost

$$\sum_{l=1}^{L} 2^l \mathcal{O}\big(\big(N/2^l\big(1 + C_{sp}\big) + k\big)k^2\big). \tag{16}$$

For fixed rank, this expression results in an asymptotic complexity of $\mathcal{O}(k^2 N \log N)$, where $L \sim \log N$. However, for high-frequency Helmholtz problems, k is a function of cluster size. In the absolute worst case for surface integral equations, $k \sim (N/2^l)^{1/2}$, a result of k growing proportionally to electrical size. Evaluating the sum in this case yields $\mathcal{O}(N^2 \log N)$.

fNCA presents difficulties as we scale to model sizes above one million unknowns. This arises as a result of the large dimensionality of $A^{(l)}$, which contains $(N/2^l)C_{sp} + k$ columns, requiring a significant amount of workspace to compute an ACA representation. The high memory requirement leads to challenges for parallelization, as multiple processes require duplications of that workspace and buffer space to store basis data. One potential solution is to develop an external-memory (out of core) ACA implementation, but the amount of I/O required would likely be intractable.

Perhaps a more concerning issue is the behavior of the sparsity constant at large electrical sizes. C_{sp} is only independent of electrical size for low-frequency problems, i.e., those where only (2) applies. This is due to the fact that the minimum admissibility distance is proportional only to the size of the cluster, so taking smaller subdivisions does not increase the number of admissible blocks relative to the parent level at saturation. For high-frequency problems where (1) also applies, the admissibility distance is proportional to $\kappa(\mathtt{diam}(t))^2$, indicating that larger electrical sizes will lead to more admissible blocks and higher values of C_{sp}, thus adding additional N dependence to (16). However, for a sufficiently small choice of η, we should recover electrical size invariance of C_{sp} by bringing the smallest clusters into a low-frequency condition $\kappa\mathtt{diam}(t) \leq 1.0$.

The small clusters which result from constraining C_{sp} will have far fields with very large relative angular breadth. As a result, k will exhibit near worst case growth with cluster size, leading to a super-quadratic asymptotic complexity for fNCA. This motivates a distinction between the cluster basis rank k and the rank of the low-rank approximation of individual admissible blocks, $\tilde{k}$. When considering the coupling between two admissible clusters, the angular breadth of one with respect to another is likely to be significantly smaller than that of their overall far fields (see Fig. 2). Hence, we expect $\tilde{k}$ to exhibit significantly less high-frequency rank growth compared with k.

The bulk of computational cost for NCA algorithms is the computation of matrix elements during ACA construction of low-rank approximations, owing to numerical integration

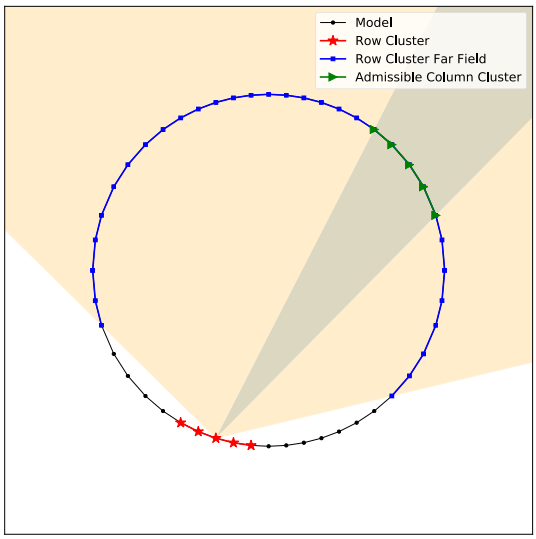


Fig. 2. Example 2-D model of an infinite cylinder cross section. A nominal row cluster and far-field, as well as an admissible column cluster are highlighted. The angular breadth of the column cluster with respect to the row cluster is shown as significantly smaller than that of its far-field.

required to construct successive rank-one approximations. rNCA exploits this fact to achieve superior practical runtime complexity, while maintaining the same asymptotic complexity as fNCA. This is achieved by performing ACA on individual admissible sub-matrices within the far-field coupling matrix, and recompressing the result into a cluster basis representation. As a result, the run time consumed by numerical integration becomes dependent on $\tilde{k}$ rather than k.

Consider the level-specific far-field coupling matrix for row cluster t at level l

$$Z_{t,F_t^{(l)}} \equiv \left(Z_{ts_1}\ Z_{ts_2}\ \cdots\ Z_{ts_x}\right) \tag{17}$$

where $\{s_1, s_2, \ldots, s_x\}$ are clusters at level l which are admissible with t. For each admissible cluster s_i, we compute the ACA representation of Z_{ts_i}, and extract the column pivots σ_i. For a single cluster t, this method requires at worst C_{sp} evaluations of ACA on square matrices of size $N/2^l$, leading to a total cost that can be modeled by

$$\sum_{l=1}^{L} 2^l \mathcal{O}\left(\left(2C_{sp}N/2^l\right)\tilde{k}^2\right). \tag{18}$$

This leads to a $\tilde{k}^2 N \log N$ practical complexity.

Assuming cluster bases for parent clusters have already been constructed, we start from the precomputed rank-k set of parent far-field pivots $\bar{\tau}_p$, and append columns to construct an approximation to $A^{(l)}(t)$

$$\tilde{A}^{(l)}(t) = \left(Z_{t\bar{\tau}_p}\ Z_{t\sigma_1}\ Z_{t\sigma_2}\ \cdots\ Z_{t\sigma_x}\right). \tag{19}$$

At the end of the procedure, $\tilde{A}^{(l)}(t)$ is recompressed using ACA to select proper pivots τ and $\bar{\tau}$ for cluster t. The cost of the recompression stage is dependent on k rather than $\tilde{k}$, because this computation constructs the final cluster basis. The overall time complexity for rNCA is thus augmented by a recompression term

$$\sum_{l=1}^{L} 2^l \mathcal{O}\left(\left(N/2^l + C_{sp}\tilde{k}\right)k^2\right). \tag{20}$$

TABLE I
OBJECT ELECTRICAL SIZE AND NUMBER OF RWG UNKNOWNS ASSOCIATED WITH NASA ALMOND SURFACE MESHES AT GIVEN FREQUENCIES

f (GHz)	Length (λ)	N
13.5	11.25	117,813
30.0	25.00	575,196
40.0	33.34	1,018,680
80.0	66.67	4,049,070
120.0	100	9,083,562
240.0	200	36,113,784

This means that rNCA ultimately has the same asymptotic scaling as fNCA. However, all required matrix elements have been precomputed in the previous ACA approximations of individual sub-matrices, so numerical integration is not required here. We show in Section III-C that while the asymptotic scaling is the same, the term which dominates in asymptotics is two orders of magnitude smaller than the numerical integration term for practical problem sizes.

Furthermore, rNCA strongly mitigates the need to take ACA algorithms out of core, as the large coupling matrices required by fNCA are reduced to more manageable sub-blocks. However, having a fully in-core ACA implementation does eventually limit the maximum size of those sub-blocks, and therefore the highest level at which clusters may be considered for admissibility. Additionally, the lower memory cost associated with computing single sub-blocks enables more cluster-basis subtrees to be constructed simultaneously in parallel. Intrasubtree parallelism is also enabled, as ACA can be run on individual sub-blocks independently. Furthermore, at any time during the construction of a cluster basis, $\tilde{A}^{(l)}(t)$ can be recompressed to reduce memory consumption, in addition to the final recompression used to construct the ultimate cluster basis representation. To enable large problem sizes, our implementation employs process parallelism with MPI to construct subtrees simultaneously, and it features an out-of-core implementation which requires only those subtrees and matrix blocks which are immediately being computed to be in main memory during execution.

III. NUMERICAL RESULTS

To showcase the NCA methods presented in this article for electrically-large problems, we compute MoM impedance matrix and radar cross sections (RCSs) for the NASA Almond [22] and a sphere, which are both treated as perfect electrical conductors (PEC). For each model, we choose several frequencies and create surface meshes with average mesh width $h = \lambda/20$. Table I lists six frequencies for the NASA Almond, along with the corresponding length in wavelengths and number of Rao–Wilton–Glisson (RWG) unknowns, and Table II lists the corresponding quantities for the sphere geometry. All of these models can be considered electrically large, given that the smallest spatial extents are about an order of magnitude greater than one wavelength.

All calculations were performed on single cluster nodes with 32 cores with two threads each. Each processor has a frequency of 2900 MHz. Process parallelism was employed

TABLE II
OBJECT ELECTRICAL SIZE AND NUMBER OF RWG UNKNOWNS ASSOCIATED WITH SPHERE SURFACE MESHES AT GIVEN FREQUENCIES

f (GHz)	Diameter (λ)	N
1.364	9.10	326,760
3.000	20.00	1,834,473
3.750	25.00	2,480,163
5.100	34.00	3,726,588
7.000	46.67	6,619,977

with MPI, using 32 individual processes for each run. Minimal communication between processes was required as cluster basis calculations can be performed independently within subtrees, and all processes had access to the same memory and disk space. No multithreading was implemented in this work.

A. Tunability of $\mathcal{H}^2$-Matrix Methods

The accuracy of both $\mathcal{H}$- and $\mathcal{H}^2$-matrix methods should be controllable via tuning of two parameters: 1) the ACA convergence threshold ε and 2) the admissibility parameter η.

First, we consider the convergence threshold for ACA, which we call ε_{NCA} and ε_{ACA} for NCA and $\mathcal{H}$-matrix ACA calculations, respectively. We have found find that ε_{NCA} should be at least as small as ε_{ACA}, consistent with the observation that $\mathcal{H}^2$-matrix representations require multiple pivoting approximations rather than the single pivoting approximation required to construct a rank-$\tilde{k}$ approximation in $\mathcal{H}$-matrix ACA.

Since $\mathcal{H}$-matrices approximate chunks of the geometry locally, matrix blocks involving regions with complicated geometries can be represented with a relatively low-compression approximation without significantly altering the overall compression for the impedance matrix. In contrast, owing to their global data sparsity, $\mathcal{H}^2$-matrices have no mechanism to handle complicated geometries locally and thus require stronger convergence tolerance to explain these features.

In addition to the convergence threshold, we also note the importance of choosing an adequate admissibility parameter η, which sets the strictness of the admissibility condition [see (1)], and thus controls the coarseness of the hierarchical partition. Owing again to the locality of data sparsity, the accuracy of $\mathcal{H}$-matrices is less sensitive to changes in η, so it is common practice to use a very high value corresponding to a coarse partition and greater compression.

In Fig. 3, we see that the η sensitivity of compression and error is much stronger for $\mathcal{H}^2$-matrix storage than for $\mathcal{H}$-matrices. These results show the rapid trade-off between compression/accuracy and admissibility for values of η between 1 and 10, with changes becoming less dramatic for higher values. Matrix storage seems to bottom out around $\eta = 12$, possibly due to the emergent frequency dependence of C_{sp} for electrically large clusters [15]. Despite this, error continues to gradually increase, so clearly for this specific example we should choose $\eta \leq 12$. We find that for many surface problems, $\eta = 10$ provides a good balance. However, this is not always the case; a similar study conducted for spheres reveals that sufficient accuracy may be achieved at $\eta = 20$.

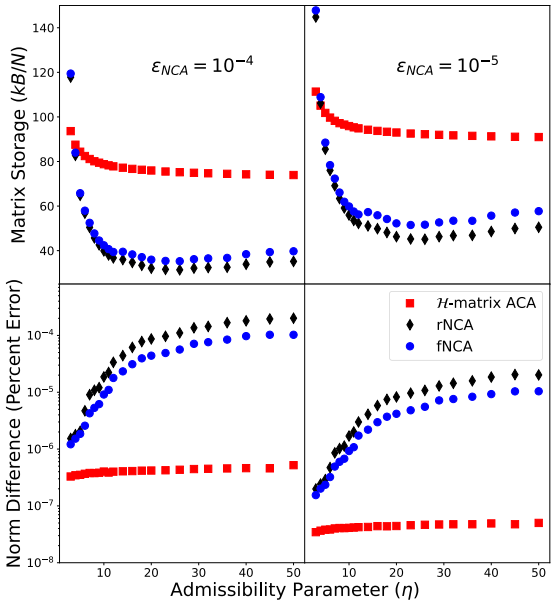


Fig. 3. Disk storage and norm difference error dependence on admissibility parameter for the NASA Almond at 13.5 GHz. Results for rNCA, fNCA, and $\mathcal{H}$-matrix ACA are shown at $\varepsilon_{\text{NCA}} = \varepsilon_{\text{ACA}} = 10^{-4}$ (left column) and 10^{-5} (right column).

We also note that changing ε_{NCA} or ε_{ACA} results in an overall shifting of the curves for error and storage for all methods. In practice, calculations should be checked for convergence using multiple values of ε_{NCA} to demonstrate an unchanging RCS. We find that $\varepsilon_{\text{NCA}} = 10^{-4}$ achieves converged RCS results for pipes and spheres, but for electrically large Almonds additional accuracy is needed to describe the sharp features around the tip. We have found that $\varepsilon_{\text{NCA}} = 10^{-5}$ was sufficient for those models.

We see also from these figures that the fNCA and rNCA methods have very similar performance trends, with fNCA providing slightly better accuracy but slightly worse compression. Compared with $\mathcal{H}^2$-matrix methods, $\mathcal{H}$-matrix methods generate significantly more accurate approximations of the overall impedance matrix, at the cost of significantly higher storage requirements. However, in Section III-B, we will demonstrate that even with less accurate approximations of impedance matrices, $\mathcal{H}^2$-matrix methods generate RCS predictions consistent with $\mathcal{H}$-matrix methods.

B. Scattering Observables for Electrically Large Models

To validate the NCA method presented here, we implemented the $\mathcal{H}^2$-matrix LU decomposition devised in [23]. We use a very high truncation tolerance to minimize errors introduced in the basis-update phase of the LU decomposition. While we have demonstrated that the NCA fill method is able to populate impedance matrices for problems up to 36 million unknowns, our current LU implementation is limited to a few million unknowns, and thus RCS results are provided only for these smaller cases.

We validate rNCA and fNCA against the Mie series solution for a PEC sphere scatterer. Fig. 4 shows the azimuthal and polar angle polarizations of the bistatic RCS of a sphere of

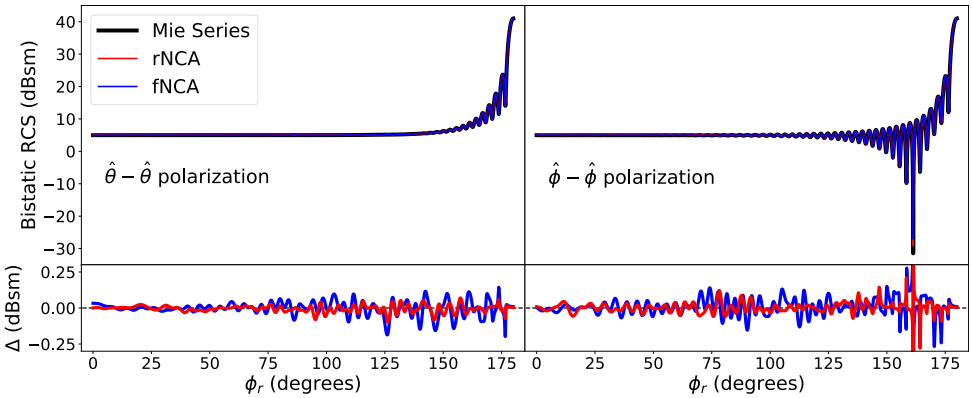


Fig. 4. Computed bistatic RCS for a 20λ diameter PEC sphere employing rNCA and fNCA, compared with Mie series analytic result. (Bottom) Residuals are plotted.

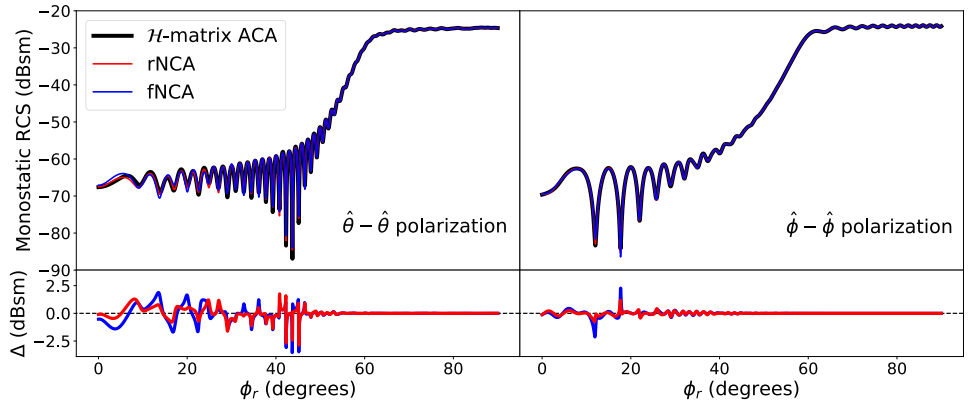


Fig. 5. Computed monostatic RCS for the NASA Almond at 40 GHz employing rNCA and fNCA, compared with $\mathcal{H}$-matrix ACA. (Bottom) Residuals are plotted.

diameter 20λ. Both NCA methods are computed with $\eta_{\text{NCA}} = 20$ and $\varepsilon_{\text{NCA}} = 10^{-4}$. Results are very consistent with the analytic Mie series, with residuals on the order of 0.1 dBsm. The largest residuals appear at nulls in the RCS. We note that residuals are roughly similar for both rNCA and fNCA.

To test the robustness of the method against multiple geometries, we consider the relatively more sophisticated NASA Almond, with a large dynamic range in the monostatic RCS. Fig. 5 shows azimuthal and polar angle polarizations for the monostatic RCS of an Almond at 40 GHz (33.34λ length). Here we cannot validate against analytic expressions, so we have computed RCS curves using $\mathcal{H}$-matrix ACA. We use $\eta_{\text{NCA}} = 10$ and $\varepsilon_{\text{NCA}} = 10^{-5}$, as we have found that the more relaxed parameters used for the sphere are not adequate to describe the low RCS region resulting from scattering off the tip of the Almond. Again, both rNCA and fNCA show good agreement with the reference RCS curve. Residuals are observed to be much larger for low RCS values, with some minor but noticeable deviations evident. Again, the largest residuals are exhibited at the nulls of the RCS, and both NCA methods feature very similar residual profiles. However, as these results are compared to those computed with an $\mathcal{H}$-matrix ACA solution and not an analytic or uncompressed solution, we can only conclude that there are minor differences between $\mathcal{H}$-matrix and $\mathcal{H}^2$-matrix results.

C. Run-Time Scaling

Despite significant rank suppression, high-frequency rank growth still remains in the absence of additional directional subdivision of the far fields. Figs. 6 and 7 show the growth of the cluster basis ranks k, and the submatrix ranks $\tilde{k}$ as a function of cluster size for the sphere and Almond geometries, respectively. The sphere is a near-worst case scenario for high-frequency rank growth, as spherical symmetry allows for far fields which frequently violate directional admissibility dramatically (see Fig. 2). This fact is evidenced in Fig. 6, where both k and $\tilde{k}$ grow as $n^{1/2}$, where n is the individual cluster size. We note, however, that $\tilde{k}$ exhibits relative boundedness for clusters smaller than 10^4 elements. In this case, both fNCA and rNCA will exhibit practical scaling like $N^2 \log N$, as will $\mathcal{H}$-matrix methods.

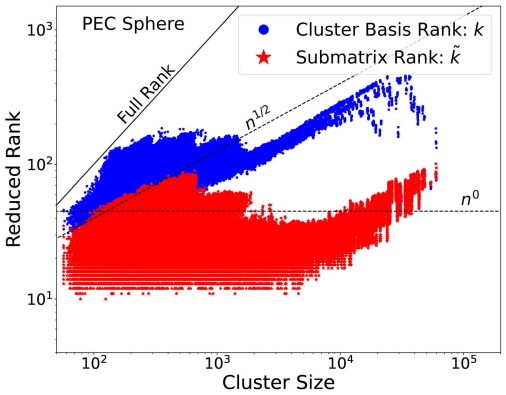


Fig. 6. Rank dependence on cluster size for the cluster basis rank k and individual submatrix ACA ranks $\tilde{k}$. Aggregated data for PEC spheres at 3, 4, and 5 GHz are shown. All calculations are performed at $\varepsilon_{\text{NCA}} = 10^{-4}$ and $\eta_{\text{NCA}} = 20$. Here, n indicates the size of individual clusters.

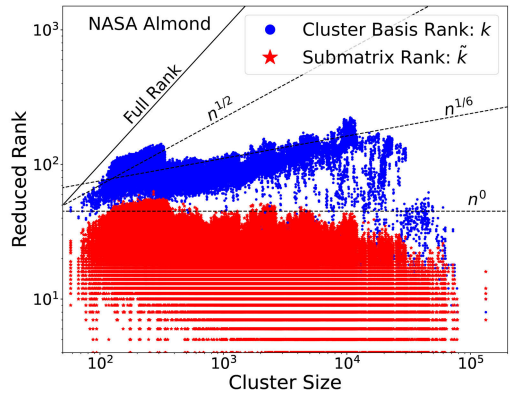


Fig. 7. Rank dependence on cluster size for the cluster basis rank k and individual submatrix ACA ranks $\tilde{k}$. Aggregated data for the NASA Almond at 40, 80, and 120 GHz are shown. All calculations are performed at $\varepsilon_{\text{NCA}} = 10^{-5}$ and $\eta_{\text{NCA}} = 10$. Here, n indicates the size of individual clusters.

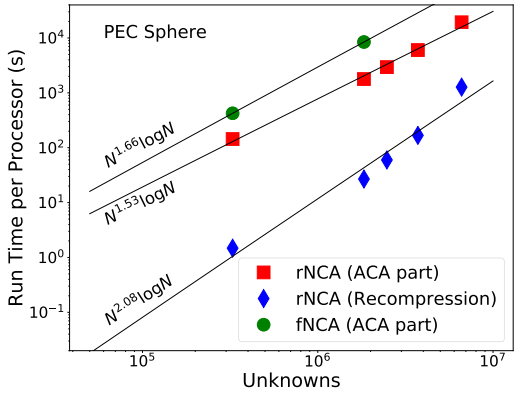


Fig. 8. Run times per process for the cluster basis computation of NCA calculations of a PEC sphere between 1 and 7 GHz. rNCA calculations are broken out into ACA and recompression parts. Dense and coupling matrix fill times are the same for rNCA and fNCA. All calculations are performed at $\varepsilon_{\text{NCA}} = 10^{-4}$ and $\eta_{\text{NCA}} = 20$.

The Almond has much more favorable rank-growth, with k exhibiting $n^{1/6}$ scaling and remarkably, $\tilde{k}$ shows complete boundedness. With these growth rates, fNCA scales as $N^{4/3}\log N$, while the practical scaling of rNCA is $N\log N$. Again, rNCA tracks with $\mathcal{H}$-matrix methods for run-time scaling for similar problems.

In Figs. 8 and 9, we plot the run times for the ACA and recompression part of rNCA and the ACA part of fNCA. Here, we compute numerical practical complexities by fitting our results to the theoretical predictions given in (16), (18), and (20). In doing so, we demonstrate the conformity of our results to the theoretical complexities expected due to the fNCA and rNCA formulations. We note that these are not asymptotic complexities but rather practical complexities for problems with large electrical sizes. It is not certain that the practical scaling relationships found here will continue to hold for significantly larger problems. To compile these results, we employed a highly optimized RWG fill function and ACA implementation from an existing CEM engine.

As predicted, Fig. 8 indicates relatively poor run time scaling across the board for the PEC sphere, with the dominant component of the run time from the ACA fill stage. Although the recompression part of rNCA exhibits super-quadratic scaling, it is multiple orders of magnitude faster than the ACA part,

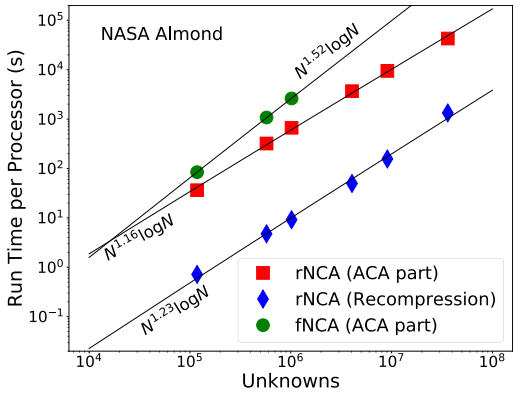


Fig. 9. Run times per process for the cluster basis computation of NCA calculations of the NASA Almond between 13.5 and 240 GHz. rNCA calculations are broken out into ACA and recompression parts. Dense and coupling matrix fill times are the same for rNCA and fNCA. All calculations are performed at $\varepsilon_{\text{NCA}} = 10^{-5}$ and $\eta_{\text{NCA}} = 10$.

which exhibits a more manageable $N^{1.53}\log N$ complexity. While the recompression scaling is daunting, this term will not become the dominant component of the run time until around two billion unknowns if the measured scaling behavior holds for larger problems. We also note that less pathological models will have significantly higher crossover points. fNCA exhibits $N^{1.66}\log N$ complexity. While this is preferable to the super-quadratic scaling of the recompression part of rNCA, the overall cost and practical scaling of rNCA is superior for problems with millions to tens of millions of unknowns, and perhaps larger if scaling relationships hold. We were not able to compute fNCA results for larger spheres due to memory limitations of our implementation.

In contrast to the sphere results, electrically large Almonds have excellent scaling properties for both rNCA and fNCA (see Fig. 9). Here, we see that both the ACA and recompression parts of the rNCA fill have time complexities near $N^{1.2}\log N$. Again, the recompression part is multiple orders of magnitude faster than the ACA part. fNCA shows markedly worse scaling of $N^{1.52}\log N$. Some part of the discrepancy between fNCA and rNCA may be due to the need to compute noncontiguous rows of the entire far-field coupling matrix in the former method, which leads to complications in looking up basis functions and managing workspace.

It is important to note that these scaling results only account for the filling of the $\mathcal{H}^2$-matrix. A full direct solver approach to MoM also includes an LU-factorization and solve step. Development of a performant and scalable factorization is outside of the scope of this current work, but to enable direct solution of electrically very-large problems, it is critical that any associated factorization and solvers are eventually demonstrated to exhibit similar near-linear runtime scalings in these regimes.

D. Memory Scaling

Figs. 10 and 11 show the scaling of the matrix storage requirement for rNCA as a function of unknowns for the PEC sphere and NASA Almond, respectively. Results are shown in kilobytes per unknown. We chose this normalization because all methods should have a common factor of N in their storage complexity; this normalization isolates the marginal differences in scaling between methods. For reference, we have

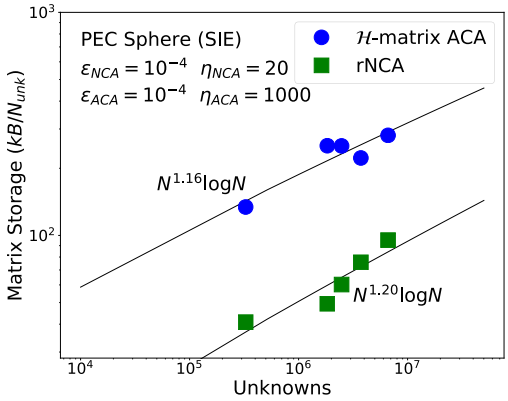


Fig. 10. Required memory for storage of impedance matrix for a PEC sphere between 1 and 7 GHz. Results are given for $\mathcal{H}$-matrix ACA and the rNCA fill methods. $\varepsilon_{\text{NCA}} = 10^{-4}$ denotes the ACA convergence tolerance used in the rNCA method. $\eta_{\text{NCA}} = 20$. $\mathcal{H}$-matrix ACA is computed at $\varepsilon_{\text{ACA}} = 10^{-4}$ and $\eta_{\text{ACA}} = 1000$.

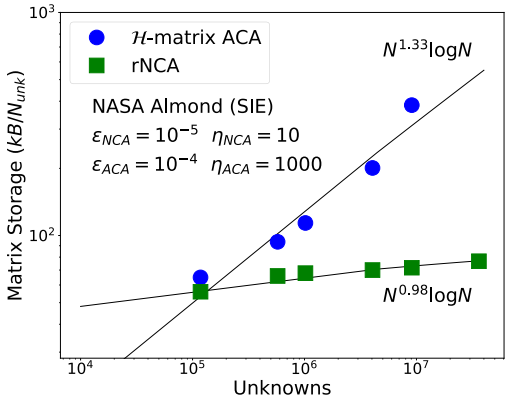


Fig. 11. Required memory for storage of impedance matrix for a NASA Almond between 13.5 and 240 GHz. Results are given for $\mathcal{H}$-matrix ACA and the rNCA fill methods. $\varepsilon_{\text{NCA}} = 10^{-5}$ denotes the ACA convergence tolerance used in the rNCA method. $\eta_{\text{NCA}} = 10$. $\mathcal{H}$-matrix ACA is computed at $\varepsilon_{\text{ACA}} = 10^{-4}$ and $\eta_{\text{ACA}} = 1000$.

included results for $\mathcal{H}$-matrix storage computed using ACA with subsequent reduced SVD to achieve near optimal compression. Because $\mathcal{H}$-matrices are more robust with respect to partitioning and error tolerance, we have used an extremely relaxed scaling factor, $\eta_{\text{ACA}} = 1000$, compared with a comparably strict factor $\eta_{\text{NCA}} = 10$–20 for the $\mathcal{H}^2$-matrix method. Furthermore, the NCA method employs the parabolic admissibility condition [see (1)], while $\mathcal{H}$-matrix ACA employs the standard condition [see (2)]. We have observed that $\mathcal{H}$-matrix methods produce excellent reproductions of impedance matrices even with these excessively liberal partitions, so this should be the standard that $\mathcal{H}^2$-matrix methods need to surpass, regardless of the relative strictness of their admissibility test. Finally, we employed higher tolerances for the NCA method where necessary to properly describe features. For $\mathcal{H}$-matrix ACA, we only consider the lower tolerance $\varepsilon_{\text{ACA}} = 10^{-4}$. This is again because $\mathcal{H}^2$-matrices are subject to significantly more error than $\mathcal{H}$-matrices due to the global approximations of matrix sparsity. fNCA results are not included, as they track very closely with rNCA results, and cannot be calculated for models as large as these due to the absence of an efficient out-of-core implementation.

For the sphere model, we observe that rNCA actually exhibits worse memory scaling than $\mathcal{H}$-matrix ACA, by a very slight margin. Despite this, the storage requirement for impedance matrices computed with rNCA is about an order of magnitude smaller, and the crossover point for these curves is unreachable (of order 10^{20} unknowns). We note that the scaling becomes significantly worse for smaller η_{NCA}. It is not surprising that the sphere exhibits poor memory scaling, given its susceptibility to high-frequency rank-growth, but we also note that there is a lot of variation in these data points, and further assessment may be required.

Results for scaling of the storage requirement are extremely impressive for the Almond meshes. We see an improvement in storage from $\mathcal{H}$-matrix ACA for all meshes considered. Most importantly, the scaling with number of unknowns has reduced from $N^{1.33}\log N$ with $\mathcal{H}$-matrix/ACA to $N^{0.98\log N}$ with $\mathcal{H}^2$-matrix/rNCA. This is a dramatic improvement that enables access to problems of significantly larger sizes. For example, if these scalings hold, the expected storage requirement for a 100 million unknown Almond would be 72.7 TB for $\mathcal{H}$-matrix ACA, while only 7.4 TB for the rNCA method. For one billion unknowns, the contrast is immense, with $\mathcal{H}$-matrix ACA requiring 1.7 PB, while the rNCA would require only a mere 79.9 TB. The factor of 22 reduction in matrix size will lead to impressive gains in overall run time (including factorization and solve steps) due to reduced I/O requirements.

If we find a degradation in accuracy for larger problems, ϵ_{NCA} may be tightened, which should ultimately result in an upward shift of the memory scaling curve, without significant change to the slope. This will increase the crossover point where $\mathcal{H}^2$-matrix methods become useful, but the overall memory scaling will still be superior to $\mathcal{H}$-matrix methods.

IV. Conclusion

In this effort, we have introduced rNCA which was designed to alleviate run time challenges associated with electrically large Helmholtz problems. We have noted that this method produces results with comparable accuracy and storage requirement to fNCA as well as improved run times and practical run time scaling. We have analyzed the rank growth behavior of spherical and Almond surface meshes and found that with a parabolic admissibility condition, rank growth is extremely manageable for the Almond geometry but scales as $n^{1/2}$ for spheres. Thus, it seems that $\mathcal{H}^2$-matrix methods are best suited to geometries which naturally obey directional admissibility.

We have demonstrated the viability of the rNCA to produce accurate impedance matrices up to at least a few million unknowns and have proven that it can efficiently achieve extremely sparse representations of problems up to 36 million unknowns which requires only a few terabytes of disk space for storage. Our storage scaling estimates indicate that impedance matrices for electrically large problems of up to one billion unknowns can be generated with fewer than 100 TB of disk space.

Acknowledgment

The authors would like to thank Sarah Longstreth, Marianne Spurrier, and Chris Plechaty for making this effort possible.

REFERENCES

[1] W. Hackbusch, "A sparse matrix arithmetic based on H-matrices. Part I: Introduction to H-matrices," *Computing*, vol. 62, no. 2, pp. 89–108, 1999.

[2] L. Grasedyck and W. Hackbusch, "Construction and arithmetics of H-matrices," *Computing*, vol. 70, no. 4, pp. 295–334, Aug. 2003.

[3] L. Banjai and W. Hackbusch, "H-and H^2-matrices for low and high frequency Helmholtz equation," Max Planck Inst. Math. Sci., Leipzig, Germany, Tech. Rep. 17/2005, 2005.

[4] M. Bebendorf, "Approximation of boundary element matrices," *Numerische Math.*, vol. 86, no. 4, pp. 565–589, Jun. 2000.

[5] M. Bebendorf and S. Rjasanow, "Adaptive low-rank approximation of collocation matrices," *Computing*, vol. 70, no. 1, pp. 1–24, 2003.

[6] W. Chai and D. Jiao, "An H-matrix-based method for reducing the complexity of integral-equation-based solutions of electromagnetic problems," in *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, Jul. 2008, pp. 1–4.

[7] O. Ergul and L. Gurel, *The Multilevel Fast Multipole Algorithm (MLFMA) for Solving Large-Scale Computational Electromagnetics Problems*. Hoboken, NJ, USA: Wiley, 2014.

[8] W. C. Chew, T. J. Cui, and J. M. Song, "A FAFFA-MLFMA algorithm for electromagnetic scattering," *IEEE Trans. Antennas Propag.*, vol. 50, no. 11, pp. 1641–1649, Nov. 2002.

[9] M. Bebendorf, C. Kuske, and R. Venn, "Wideband nested cross approximation for Helmholtz problems," *Numerische Mathematik*, vol. 130, no. 1, pp. 1–34, 2015.

[10] B. Engquist and L. Ying, "Fast directional multilevel algorithms for oscillatory kernels," *SIAM J. Sci. Comput.*, vol. 29, no. 4, pp. 1710–1737, Jan. 2007.

[11] V. Rokhlin, "Diagonal forms of translation operators for the Helmholtz equation in three dimensions," *Appl. Comput. Harmon. Anal.*, vol. 1, no. 1, pp. 82–93, 1993.

[12] H. Cheng et al., "A wideband fast multipole method for the Helmholtz equation in three dimensions," *J. Comput. Phys.*, vol. 216, no. 1, pp. 300–325, Jul. 2006.

[13] S. Börm and W. Hackbusch, "Approximation of boundary element operators by adaptive H^2-matrices," 2003.

[14] S. Börm, "Data-sparse approximation of non-local operators by H^2-matrices," *Linear Algebra Appl.*, vol. 422, nos. 2–3, pp. 380–403, Apr. 2007.

[15] S. Börm, "Directional H^2-matrix compression for high-frequency problems," *Numer. Linear Algebra Appl.*, vol. 24, no. 6, Dec. 2017, Art. no. e2112.

[16] B. Engquist and L. Ying, "A fast directional algorithm for high frequency acoustic scattering in two dimensions," *Commun. Math. Sci.*, vol. 7, no. 2, pp. 327–345, 2009.

[17] M. Bebendorf and R. Venn, "Constructing nested bases approximations from the entries of non-local operators," *Numer. Math.*, vol. 121, no. 4, pp. 609–635, Aug. 2012.

[18] A. Yu Mikhalev and I. V. Oseledets, "Iterative representing set selection for nested cross approximation," 2013, *arXiv:1309.1773*.

[19] S. Börm and S. Christophersen, "Approximation of integral operators by green quadrature and nested cross approximation," *Numerische Math.*, vol. 133, no. 3, pp. 409–442, Jul. 2016.

[20] Y. Zhao, D. Jiao, and J. Mao, "Fast nested cross approximation algorithm for solving large-scale electromagnetic problems," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 8, pp. 3271–3283, Aug. 2019.

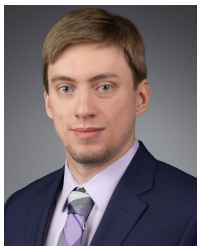
[21] J. Ostrowski, Z. Andjelic, M. Bebendorf, B. Cranganu-Cretu, and J. Smajic, "Fast BEM-solution of Laplace problems with H-matrices and ACA," *IEEE Trans. Magn.*, vol. 42, no. 4, pp. 627–630, Apr. 2006.

[22] A. C. Woo, H. T. G. Wang, M. J. Schuh, and M. L. Sanders, "EM programmer's notebook-benchmark radar targets for the validation of computational electromagnetics programs," *IEEE Trans. Antennas Propag.*, vol. 35, no. 1, pp. 84–89, Feb. 1993.

[23] M. Ma and D. Jiao, "Accuracy directly controlled fast direct solution of general H^2-matrices and its application to solving electrodynamic volume integral equations," *IEEE Trans. Microw. Theory Techn.*, vol. 66, no. 1, pp. 35–48, Jan. 2017.

Nathan M. Parzuchowski (Member, IEEE) received the Ph.D. degree in nuclear physics from Michigan State University, East Lansing, MI, USA, in 2017. His graduate thesis was focused on novel methods for the computation of nuclear excited state structure and transitions starting from first-principles interactions.

He has been a Member of the Research Staff at Riverside Research, Beavercreek, OH, USA, since 2018. His current research interests are in the method development for electromagnetics and imaging spectroscopy.



Brenton Hall received the Bachelor of Arts degree in computational mathematics with minor in physics from Asbury University, Wilmore, KY, USA, in 2015, and the Master of Mathematical Sciences degree (focusing on computational science) from The Ohio State University, Columbus, OH, USA, in 2017.

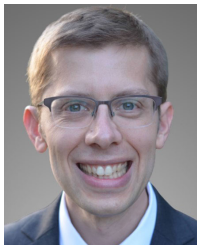
Since 2017, he has been working on multiple projects in the fields of navigation and electromagnetics with Riverside Research, Beavercreek, OH, USA.

Isroel M. Mandel (Member, IEEE) received the Ph.D. degree from the Graduate Center, City University of New York (CUNY), New York, NY, USA, in 2015, where he has focused on investigating light sorting electromagnetic metamaterials and metasurfaces.

He has been a Member of the Research Staff with Riverside Research, New York, since 2014, researching HPC-grade integral equation computation codes and hierarchical matrix arithmetic for accelerated numerical linear algebra.

Ian Holloway received the B.S. degree in physics with minors in mathematics and computer science from Cedarville University, Cedarville, OH, USA, in 2016, and the Ph.D. degree in interdisciplinary applied science and mathematics from Wright State University, Fairborn, OH, USA, in 2019, with a focus on modeling supersonic compressible and magnetohydrodynamic flows past cones.

He has been a Member of the Research Staff with Riverside Research, Beavercreek, OH, USA, since 2020. His research interests include development of numerical and statistical methods for physics and engineering applications.



Eli Lansey (Member, IEEE) received the Ph.D. degree in physics from the City University of New York, New York, NY, USA, in 2012, with his thesis focused on theoretical and numerical approaches for modeling metamaterials.

Since graduating, he has been with Riverside Research, New York. His current research interests include method development for electromagnetics and radar imaging.