Communication

Error Control of Multiple-Precision MLFMA

Mert Kalfa, Özgür Ergül, and Vakur B. Ertürk

***Abstract*— We introduce and demonstrate a new error control scheme for the computation of far-zone interactions in the multilevel fast multipole algorithm when implemented within a multiple-precision arithmetic framework. The proposed scheme provides the optimum truncation numbers as well as the machine precisions given the desired relative error thresholds and the box sizes for the translation operator at all frequencies. In other words, unlike the previous error control schemes which are valid only for high-frequency problems, the proposed scheme can be used to control the error across both low- and high-frequency problems. Optimum truncation numbers and machine precisions are calculated for a wide range of box sizes and desired relative error thresholds with the proposed error control scheme. The results are compared with the previously available methods and numerical surveys.**

***Index Terms*— Diagonalization, error analysis, fast multipole method (FMM), low-frequency breakdown, multiple-precision arithmetic (MPA).**

I. Introduction

The fast multipole method (FMM) has been named as one of the top 10 algorithms of the 20th century by the Society of Industrial and Applied Mathematics [1]. The multilevel fast multipole algorithm (MLFMA) that is an extension of FMM is able to achieve $O(N \log N)$ complexity for N unknowns, enabling the solution of extremely large electromagnetic problems compared with $O(N^2)$ for a Krylov-subspace iterative algorithm applied on full matrices. This increase in efficiency is due to the ability to compute the interactions between basis and testing functions in a group-by-group manner, which is made possible by Gegenbauer's addition theorem and the diagonalization of the translation operator [2], [3]. The drawback of the diagonalized form is that it includes an infinite summation over spherical harmonics involving Hankel functions which become numerically unstable as the truncation number (order) increases due to limited machine precision. The numerical stability problem of the translation operator is also the main culprit behind the well-known low-frequency breakdown problem [4], which makes selecting the truncation number an important part of the error control of MLFMA.

There have been several classical papers about the error control of MLFMA, and most of them focus on the translation operator and its truncation. In [5]–[7], the excess bandwidth formula (EBF) is used to determine the truncation numbers. Although it is widely used in most MLFMA implementations, there are two main limitations of the EBF. First, the formula is derived using the large argument approximation of the Bessel and Hankel functions [8] of the translation operator, which makes it invalid for small box sizes (i.e., low-frequency problems). Second, the numerical stability of the Hankel function given the available machine precision is not considered. The second shortcoming is partly addressed in [9], where the accuracy lost due to the overflow of the Hankel function is considered. However, the resulting error control scheme is only limited to electrically large boxes.

In the case of electrically small boxes, there are many different studies available in the literature to treat the low-frequency breakdown that falls into mainly two categories. One popular approach is to use the multipoles explicitly [10]–[15], while another is to deform the angular integration so that the evanescent waves are considered for subwavelength interactions [16]–[18]. Methods in both categories require the solver to be implemented from the ground up while increasing complexity due to alternative expansion formulations of Green's function. A simple alternative to the treatment of the low-frequency problem is proposed in [19], where multiple-precision arithmetic (MPA) is used to handle overflowing summations when necessary. Since the EBF is not valid for low-frequency problems, Ergül and Karaosmanoğlu [19] determined the optimum truncation numbers and machine precisions by extensive numerical simulations.

In this communication, we introduce and demonstrate an error control scheme for MLFMA that is valid at all frequencies when implemented in a multiple-precision framework. The proposed scheme provides the optimum truncation numbers given the box size and the desired relative error threshold at all frequencies for the first time in the literature while yielding compatible results with the EBF at high frequencies. In addition, the proposed scheme provides the required machine precisions for each case, results of which can be used as a precursor to an efficient and robust implementation of an MPA-based MLFMA solver.

The rest of this communication is organized as follows. Section II describes the proposed error control formulation and its implementation. Section III presents the numerical results and comparisons with the existing methods and numerical surveys in the literature. Section IV presents a discussion on the implementation of MPA and the required computing resources. The conclusion is provided in Section V. An $e^{-i\omega t}$ time convention, where $\omega = 2\pi f$ and f is the operating frequency, is assumed and suppressed throughout this communication.

II. Formulation

A. Error Control Formulations in the Literature

Gegenbauer's addition theorem expands the free-space Green's function in terms of spherical harmonics as

$$\frac{\exp(ik|\vec{w}+\vec{v}|)}{4\pi|\vec{w}+\vec{v}|} = \frac{ik}{4\pi}\sum_{t=0}^{\infty}(-1)^t(2t+1)j_t(kv)h_t^{(1)}(kw)P_t(\hat{w}\cdot\hat{v}) \tag{1}$$

where j_t and $h_t^{(1)}$ are the spherical Bessel and Hankel functions of the first kind, respectively. In (1), P_t is the Legendre polynomial of order t, while $w = |\vec{w}|$ and $v = |\vec{v}|$ represent the translation and

Manuscript received November 30, 2017; revised May 23, 2018; accepted June 27, 2018. Date of publication July 9, 2018; date of current version October 4, 2018. *(Corresponding author: Mert Kalfa.)*

M. Kalfa is with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey, and also with the Aselsan Research Center, Aselsan Inc., 06370 Ankara, Turkey (e-mail: kalfa@ee.bilkent.edu.tr).

Ö. Ergül is with the Department of Electrical and Electronics Engineering, Middle East Technical University, 06800 Ankara, Turkey (e-mail: ozgur.ergul@eee.metu.edu.tr).

V. B. Ertürk is with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: vakur@ee.bilkent.edu.tr).

Color versions of one or more of the figures in this communication are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAP.2018.2854405

shift vectors, respectively. Note that (1) is only valid when $w > v$. Representing the spherical waves as integrals over the plane-wave spectrum, the diagonal form of Green's function is obtained [2] as

$$\frac{\exp(ik|\vec{w}+\vec{v}_1+\vec{v}_2|)}{4\pi|\vec{w}+\vec{v}_1+\vec{v}_2|} \approx \frac{ik}{(4\pi)^2}\int d^2\hat{k}\,\beta(\vec{k},\vec{v}_1)\alpha_\tau(\vec{k},\vec{w})\beta(\vec{k},\vec{v}_2) \tag{2}$$

where

$$\beta(\vec{k},\vec{v}) = \exp(i\vec{k}\cdot\vec{v}) \tag{3}$$

$$\alpha_\tau(\vec{k},\vec{w}) = \sum_{t=0}^{\tau}(i)^t(2t+1)h_t^{(1)}(kw)P_t(\hat{k}\cdot\hat{w}) \tag{4}$$

are the shift and the translation operators, respectively. In (4), τ is the truncation number that directly affects the accuracy of the translation operator and ultimately the accuracy of MLFMA. The truncation number also determines the number of sampling points along the θ and ϕ axes of the spherical coordinate system [20] for the angular integration in (2). For electrically large boxes, the EBF is used to determine the truncation number [5]–[7] as

$$\tau \approx ka\sqrt{3} + 2.18(d_0)^{2/3}(ka)^{1/3} \tag{5}$$

where a is the box edge length and d_0 is the desired digits of accuracy which is related to the desired relative error threshold (ϵ_d) by

$$d_0 \triangleq -\log_{10}(\epsilon_d). \tag{6}$$

In practice, the actual relative error of (2) with respect to the free-space Green's function may exceed the desired threshold (ϵ_d) due to overflow problems in the evaluation of the spherical Hankel function on a computing platform with a limited machine precision. This behavior of the Hankel function is addressed in [9], where the digits of accuracy lost due to the spherical Hankel function are modeled as

$$d_1 = \left[\frac{\tau - 2ka}{1.8(2ka)^{1/3}}\right]^{1.5} \tag{7}$$

which can be used to estimate the effective digits of accuracy given by

$$d_{\text{eff}} = d_0 - d_1. \tag{8}$$

Note that the accuracy estimate of (8) is still only valid for electrically large boxes since it is based on the large argument approximation of Hankel functions.

B. Proposed Error Control Formulation

1) Estimating the Optimum Truncation Number: The derivation for the relative error starts with Gegenbauer's addition theorem (1). When a truncation number of τ is used, the relative error ($\hat{\epsilon}$) with respect to the free-space Green's function can be found from

$$\hat{\epsilon} = kR\left|\sum_{t=\tau+1}^{\infty}(-1)^t(2t+1)j_t(kv)h_t^{(1)}(kw)P_t(\hat{w}\cdot\hat{v})\right| \tag{9}$$

where $R = |\vec{w}+\vec{v}|$. Assuming that the leading term of (9) dominates, the relative error can be approximated as

$$\hat{\epsilon} \approx kR(2\tau+3)\left|j_{\tau+1}(kv)h_{\tau+1}^{(1)}(kw)P_{\tau+1}(\hat{w}\cdot\hat{v})\right|. \tag{10}$$

Unlike the previous works on error control [5]–[7] that use the large argument approximations of the Bessel and Hankel functions in (10), we use the large order approximation [8, eq. (9.3.2)] as

$$J_t(t\operatorname{sech}\gamma_j) \approx \frac{\exp[t(\tanh\gamma_j - \gamma_j)]}{\sqrt{2\pi t\tanh\gamma_j}} \tag{11}$$

$$Y_t(t\operatorname{sech}\gamma_y) \approx -i\frac{\exp[t(\gamma_y - \tanh\gamma_y)]}{\sqrt{\frac{1}{2}\pi t\tanh\gamma_y}}. \tag{12}$$

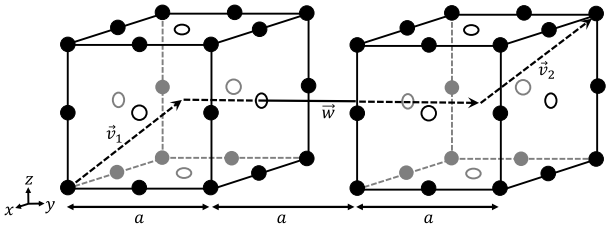

Fig. 1. Critical points for the shift vectors shown on source and observation boxes in a one-box-buffer scheme. Corners and edge centers are shown with filled circles, and face centers are shown as circles.

Substituting (11) and (12) into the spherical Bessel and Hankel functions in (10), we obtain

$$j_{\tau+1}(kv) = \sqrt{\frac{\pi}{2kv}}J_{\tau+1.5}(kv) \approx \frac{0.5\psi_j}{\sqrt{(\tau+1.5)kv\tanh\gamma_j}} \tag{13}$$

$$h_{\tau+1}^{(1)}(kw) = \sqrt{\frac{\pi}{2kw}}H_{\tau+1.5}^{(1)}(kw) \approx \frac{0.5\psi_h - i\,\psi_h^{-1}}{\sqrt{(\tau+1.5)kw\tanh\gamma_h}} \tag{14}$$

where ψ_j and ψ_h are defined as

$$\psi_j \triangleq \exp[(\tau+1.5)(\tanh\gamma_j - \gamma_j)] \tag{15}$$

$$\psi_h \triangleq \exp[(\tau+1.5)(\tanh\gamma_h - \gamma_h)] \tag{16}$$

with

$$\gamma_j = \operatorname{sech}^{-1}\left(\frac{kv}{\tau+1.5}\right) \tag{17}$$

$$\gamma_h = \operatorname{sech}^{-1}\left(\frac{kw}{\tau+1.5}\right). \tag{18}$$

Substituting (13) and (14) into (10), we obtain

$$\hat{\epsilon} \approx \frac{R}{\sqrt{wv}}\left|\frac{P_{\tau+1}(\hat{w}\cdot\hat{v})\psi_j\left(0.5\psi_h - i\,\psi_h^{-1}\right)}{\sqrt{\tanh\gamma_j\,\tanh\gamma_h}}\right| \tag{19}$$

which can be used to estimate the relative error for all possible translation and shift vectors.

To find the maximum error given a box size (a), the translation distance is taken as its minimum value (i.e., $w = 2a$ for a one-box-buffer scheme) and the total shift distance is taken as its maximum value ($v = |[a\ \ a\ \ a]^T| = a\sqrt{3}$, where T is the matrix transpose), respectively, as shown in Fig. 1.

Note that the error control schemes used in [5]–[7] assume $|P_{\tau+1}(\hat{w}\cdot\hat{v})| = 1$ as the worst case when estimating the relative error. However, the values of the argument where the absolute value of the Legendre polynomial reaches unity ($\hat{w}\cdot\hat{v} = \pm 1$) are not necessarily the points where the largest errors will occur, e.g., $\hat{w}\cdot\hat{v} = 1/\sqrt{3}$ for the worst case illustrated in Fig. 1. Moreover, the root-mean-square value of the Legendre polynomials over the range $\hat{w}\cdot\hat{v} \in [-1, 1]$ can be calculated as

$$\sqrt{\int_{-1}^{1}P_t^2(z)dz} = \sqrt{\frac{2}{2t+1}} \tag{20}$$

which can be derived using Rodrigues' formula [8, eq. (8.6.18)] and integration by parts. Equation (20) shows that assuming $|P_{\tau+1}(\hat{w}\cdot\hat{v})| = 1$ in (19) leads to overestimation of the truncation numbers, especially for larger box sizes and/or smaller desired relative errors. Moreover, even slightly overestimated truncation numbers cause a much higher required machine precision, especially for low frequencies because of the increased numerical instability of the Hankel functions for small arguments. As a result, the Legendre polynomial is kept intact in (19).

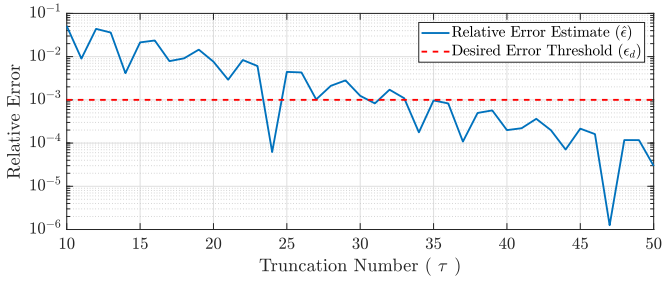

Fig. 2. Relative error estimate ($\hat{\epsilon}$) for $a = \lambda/32$, $\epsilon_d = 1e{-}3$, $\vec{w} = [0\ 2a\ 0]^T$, and $\vec{v} = \vec{v}_1 + \vec{v}_2 = [a\ a\ \ a]^T$.

To the best of our knowledge, (19) cannot be solved for τ analytically, due to the inclusion of the Legendre polynomial. Therefore, the number of harmonics for a given box size (a) and a desired relative error threshold (ϵ_d) must be found numerically. Moreover, due to the oscillatory nature of the Legendre polynomial, there are more than one solution for a given relative error threshold. This behavior is shown in Fig. 2, in which the relative error estimate as a function of the truncation number is given for an example scenario. Therefore, to provide an accurate upper bound for the relative error, we define a set of feasible truncation numbers as

$$\tilde{\tau}(a, \epsilon_d) \triangleq \{\tau \in \mathbb{Z}^+ \mid \hat{\epsilon}(a, \tau) < \epsilon_d, \hat{\epsilon}(a, \tau - 1) > \epsilon_d\}. \tag{21}$$

Then, the optimum truncation number (τ_{opt}) can simply be found as the maximum value of the feasible set as

$$\tau_{\text{opt}}(a, \epsilon_d) \triangleq \max(\tilde{\tau}(a, \epsilon_d)). \tag{22}$$

From a practical standpoint, (21) and (22) correspond to finding the zero crossings of $\epsilon_d - \hat{\epsilon}$, and then, selecting the largest τ value for which the estimated relative error is below the desired relative error. This approach has a computational complexity of $O(\tau_{\text{opt}})$ and ensures the actual relative error of the translation operator stays below the specified level.

2) Estimating the Optimum Machine Precision: After finding the optimum number of harmonics (τ_{opt}), the far-zone interactions between the boxes are computed using the diagonal form of Green's function in (2). The machine precision must be able to handle each of the individual elementary functions in (2), as well as all of the intermediate combinations (i.e., products, summations, and integrations) before the final result. Note that the required machine precision is highly dependent on the implementation (order of computation, canceling terms, and so on). An important assumption on the implementation is that the frequency scaling that comes from multiplication by k in (2) is performed after the computation is finished, i.e., the first k term in (2) is replaced by 2π while estimating the required machine precision.

To represent the worst case in terms of the required machine precision when computing (2), we define

$$G^{\text{MP}} \triangleq \frac{2\pi N_{\theta\phi}\Delta\theta\Delta\phi}{(4\pi)^2}(\tau+1)(2\tau+1)h_\tau^{(1)}(kw)P_\tau^{\text{MP}}. \tag{23}$$

In (23), P_τ^{MP} is defined to be the value of the Legendre polynomial $P_\tau(\hat{k}\cdot\hat{w})$ that requires the highest machine precision (i.e., its minimum value other than zero) as

$$P_\tau^{\text{MP}} \triangleq \left\{ \min_{\hat{k}\in\mathbb{K}^3} (|P_\tau(\hat{k}\cdot\hat{w})|) \,|\, P_\tau(\hat{k}\cdot\hat{w}) \neq 0 \right\} \tag{24}$$

where $\mathbb{K}^3$ is the set of unit vectors $\hat{k}$ that are defined by the angular sampling in the numerical evaluation of (2). The steps taken to obtain (23) from (2) are as follows. First, the unit amplitude shift operators ($\beta(\vec{k}, \vec{v})$) are omitted. Second, assuming that all terms in the truncated summation in (4) coherently adds up as the worst case, the summation is replaced with a multiplication by the number of terms, i.e., $(\tau + 1)$. Third, assuming that the integrand coherently adds up in (2), the integral is replaced with a multiplication by $N_{\theta\phi}\Delta\theta\Delta\phi$, where $\Delta\theta$ and $\Delta\phi$ are the grid sizes along the θ and ϕ axes, respectively, and $N_{\theta\phi}$ is the total number of angular samples. Note that we use Gauss–Legendre sampling along the θ-axis as given in [20] when computing (2). However, while constructing (23), we assume uniform sampling, which yields

$$\Delta\theta = \Delta\phi = \frac{\pi}{\tau+1} \tag{25}$$

$$N_{\theta\phi} = 2(\tau+1)^2. \tag{26}$$

Note that uniform integration weights follow the sample mean of Gauss–Legendre weights with a multiplicative factor of $\pi/2$ (i.e., half of the extent of θ-axis). Therefore, uniform integration weights offer a simpler and computationally tractable alternative when computing (23).

The expression given in (23) includes the multiplicative terms both greater and smaller than one, which we define as overflow-critical (G_+^{MP}) and underflow-critical terms (G_-^{MP}), respectively, as

$$G_+^{\text{MP}} = 2\pi N_{\theta\phi}(\tau+1)(2\tau+1)\max\left(0.5|\psi_h|, \left|\psi_h^{-1}\right|\right) \tag{27}$$

$$G_-^{\text{MP}} = \frac{\Delta\theta\Delta\phi}{(4\pi)^2}\frac{P_\tau^{\text{MP}}}{\sqrt{(\tau+1.5)kw\tanh\gamma_h}}. \tag{28}$$

Note the spherical Hankel function in (23) is replaced by its large order approximation in (14), and only the dominating term of the numerator of (14) is considered ($|\psi_h| \gg 1$ for large boxes and vice versa). In the worst case, the machine precision must be large enough to handle both G_+^{MP} and G_-^{MP} separately. Therefore, the required decimal digits of machine precision for computing (23) can be found as

$$\text{MP}_G = \max\left(\log_{10}\left(G_+^{\text{MP}}\right), -\log_{10}\left(G_-^{\text{MP}}\right)\right). \tag{29}$$

When determining the optimum machine precision, we must also consider the actual expected amplitude of Green's function and the desired relative error threshold as

$$\text{MP}_\epsilon \triangleq -\log_{10}(\epsilon_d) + \log_{10}(4\pi R_{\max}) + 1 \tag{30}$$

where $\epsilon_d \in (0, 1)$ and $4\pi R_{\max}$ is the denominator of the free-space Green's function with $R_{\max}$ as the maximum value of R for the given box size (for the one-box-buffer scheme, $\vec{v} = [a\ a\ \ a]^T$ and $R_{\max} = a\sqrt{11}$). The $+1$ term in (30) is added empirically for safety. Finally, the optimum machine precision (MP_{opt}) for computing (2) can be found as

$$\text{MP}_{\text{opt}} = \lceil \max(\text{MP}_G, \text{MP}_\epsilon) \rceil. \tag{31}$$

To summarize, given a box size (a) and a desired relative error threshold (ϵ_d), (19)–(22) can be used to find the optimum truncation number (τ_{opt}), while (27)–(31) can be used to find the optimum digits of machine precision (MP_{opt}). Note that, (19)–(22) and (27)–(31) can also be used to infer the achievable relative errors and the corresponding truncation numbers given the available machine precision.

III. Numerical Results

The proposed error control scheme was implemented in MATLAB, where the MPA environment was constructed using a commercially available toolbox [21]. In order to validate the proposed error control scheme, the following scenarios were investigated.

1) *Box Size (a):* 64λ to $\lambda/2048$ in base-2 logarithmic steps.
2) *Desired Relative Error:* $\epsilon_d \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

TABLE I

OPTIMUM TRUNCATION NUMBERS (τ_{opt}) AND MACHINE PRECISIONS (MP_{opt}) FOR VARIOUS DESIRED RELATIVE ERROR THRESHOLDS (ϵ_d) AND BOX SIZES (a) WHEN $\vec{w} = [0\ 2a\ 0]^T$

Error Limit→	10^{-2}		10^{-3}		10^{-4}		10^{-5}	
Box Size ↓	τ_{opt}	MP_{opt}	τ_{opt}	MP_{opt}	τ_{opt}	MP_{opt}	τ_{opt}	MP_{opt}
64λ	714	15	721	15	728	15	734	15
32λ	360	14	368	14	375	14	382	14
16λ	185	13	191	13	197	13	201	13
8λ	96	10	102	10	106	11	115	13
4λ	50	10	59	11	69	14	80	19
2λ	30	9	43	15	56	23	69	32
λ	24	13	37	22	50	34	66	50
$\lambda/2$	20	16	36	32	50	49	66	70
$\lambda/4$	20	22	34	41	50	64	66	90
$\lambda/8$	20	29	34	51	50	80	66	110
$\lambda/16$	20	35	34	62	50	95	66	131
$\lambda/32$	20	42	34	73	50	111	66	151
$\lambda/64$	20	48	34	83	50	126	66	171
$\lambda/128$	20	55	34	94	50	142	66	192
$\lambda/256$	20	61	34	105	50	157	66	212
$\lambda/512$	20	68	34	116	50	173	66	232
$\lambda/1024$	20	74	34	126	50	189	66	253
$\lambda/2048$	20	81	34	137	50	204	66	273

TABLE II

OPTIMUM TRUNCATION NUMBERS (τ_{opt}) AND MACHINE PRECISIONS (MP_{opt}) FOR VARIOUS DESIRED RELATIVE ERROR THRESHOLDS (ϵ_d) AND BOX SIZES (a) WHEN $\vec{w} = [3a\ 3a\ 3a]^T$

Error Limit→	10^{-2}		10^{-3}		10^{-4}		10^{-5}	
Box Size ↓	τ_{opt}	MP_{opt}	τ_{opt}	MP_{opt}	τ_{opt}	MP_{opt}	τ_{opt}	MP_{opt}
64λ	724	15	732	15	739	15	745	15
32λ	370	14	376	14	381	14	386	14
16λ	190	14	195	14	200	14	204	14
8λ	100	11	103	11	108	13	111	13
4λ	54	10	56	10	59	10	62	10
2λ	29	8	32	9	35	9	36	10
λ	17	8	19	8	21	9	23	10
$\lambda/2$	10	7	12	8	14	9	15	10
$\lambda/4$	7	7	8	8	10	9	11	10
$\lambda/8$	5	7	6	8	8	9	10	10
$\lambda/16$	4	7	6	9	8	11	10	13
$\lambda/32$	4	8	6	11	8	14	10	17
$\lambda/64$	4	10	6	13	8	16	10	20
$\lambda/128$	4	11	6	15	8	19	10	24
$\lambda/256$	4	13	6	18	8	22	10	27
$\lambda/512$	4	15	6	20	8	25	10	31
$\lambda/1024$	4	17	6	22	8	28	10	34
$\lambda/2048$	4	18	6	25	8	31	10	38

3) *Translation Vectors ($\vec{w}$):* $[0\ 2a\ 0]^T$ for minimum translation distance along the y-axis (see Fig. 1) and $[3a\ 3a\ 3a]^T$ for maximum translation distance for a one-box-buffer scheme.
4) *Shift Vectors ($\vec{v} = \vec{v}_1 + \vec{v}_2$):* From corners, edge centers, and face centers of the source box to those of the observation box (shown in Fig. 1)

For each scenario listed earlier, we calculated the optimum truncation numbers (τ_{opt}) and the optimum digits of machine precision (MP_{opt}) using (19)–(22) and (27)–(31), which are reported in Table I for $\vec{w} = [0\ 2a\ 0]^T$ and in Table II for $\vec{w} = [3a\ 3a\ 3a]^T$. Note that for each entry in Tables I and II, we investigated all critical shift vectors and report the largest τ_{opt} and MP_{opt} pairs.

Using Tables I and II, the actual relative errors with respect to the free-space Green's function are given in Figs. 3 and 4. As shown in Figs. 3 and 4, the truncation numbers and the machine precisions obtained from the proposed scheme keep the relative errors close to or below the desired levels for both large and small boxes. Note that some small oscillations in the actual errors can be observed as the box size increases due to the large order approximations given in (11) and (12). The large order approximation becomes slightly more erroneous as the arguments of the Bessel and Hankel functions get closer to their order for very large arguments. However, the error due to the large order approximation for asymptotically large boxes is always bounded, which can be shown analytically by comparing the large order and large argument approximations [8] of the Bessel and Hankel functions. Therefore, the proposed error control scheme can be used for arbitrarily large box sizes.

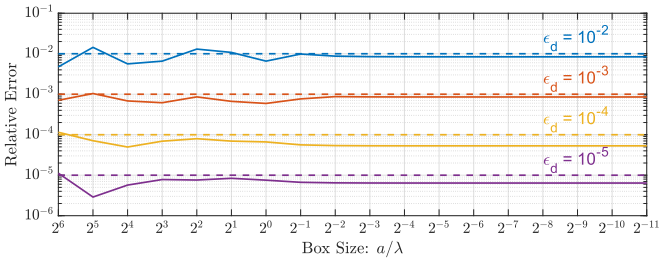

Fig. 3. Relative errors with respect to free-space Green's function when Table I is used in an MPA environment for $\vec{w} = [0\ 2a\ 0]^T$. Dashed lines represent the desired relative error thresholds.

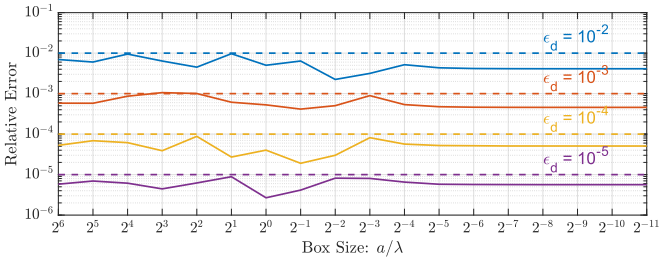

Fig. 4. Relative errors with respect to free-space Green's function when Table II is used in an MPA environment for $\vec{w} = [3a\ 3a\ 3a]^T$. Dashed lines represent the desired relative error thresholds.

An interesting observation for Tables I and II is that τ_{opt} values for a given error threshold become constant for electrically small boxes. This behavior is also observed in the harmonics of Green's function when the multipole expansion is explicitly used as in [14]. Therefore, truncation numbers only depend on the desired relative error thresholds for electrically small boxes.

Another important observation is that there is always a minimum value of MP_{opt} from where the value increases for both increasing and decreasing box sizes. For electrically small boxes, the spherical Hankel function in (2) dominates every other term and gets larger as the box size decreases asymptotically. For electrically large boxes, the terms due to angular sampling and numerical integration given in (25) and (26) dominate, which then causes an increase in MP_{opt} as the box size increases asymptotically.

When we compare Tables I and II, we observe that as the translation distance (w) increases, the τ_{opt} and MP_{opt} pairs decrease significantly for small boxes while increasing slightly for large boxes. The behavior for small boxes is again due to the spherical Hankel function dominating the other terms in (2). For very small arguments (i.e., electrically small boxes), a relatively small increase in the translation distance from Tables I and II causes a reduction of many orders of magnitude in the value of the spherical Hankel function, leading to dramatic reductions in the estimates of both τ_{opt} and MP_{opt}. On the other hand, the slight increase of MP_{opt}

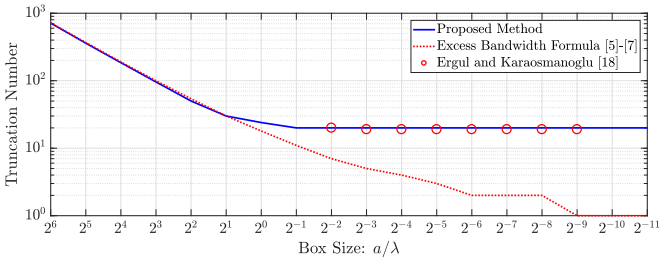

Fig. 5. Comparison of the truncation numbers found by the proposed scheme to [5]–[7] and [19] for $\epsilon_d = 10^{-2}$ and $\vec{w} = [0\ 2a\ 0]^T$.

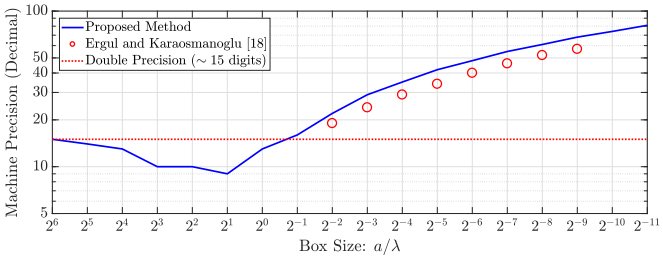

Fig. 6. Comparison of the machine precisions found by the proposed scheme to double precision and [19] for $\epsilon_d = 10^{-2}$ and $\vec{w} = [0\ 2a\ 0]^T$.

for the larger boxes is due to the first term in the right-hand side of (19) increasing for larger translation distances.

The τ_{opt} and MP_{opt} pairs obtained with our proposed scheme also agree well with the previous studies found in the literature. Fig. 5 compares the proposed scheme with the well-known EBF [5]–[7] when $\epsilon_d = 10^{-2}$ and $\vec{w} = [0\ 2a\ 0]^T$. Since the EBF is not valid for small boxes, truncation numbers found via numerical simulations in [19] are also shown in Fig. 5. The proposed scheme agrees very well with the EBF for electrically large boxes while being valid for electrically small box sizes. A similar comparison is given in Fig. 6, where the optimum machine precisions given in Table I for $\epsilon_d = 10^{-2}$ and machine precisions found numerically in [19] are shown. The proposed scheme estimates slightly larger MP_{opt} values while still following the trend found in [19]. This is expected since our method assumes the worst case in terms of the implementation for the optimum machine precision, leading to MP_{opt} estimates that are greater than or equal to experimental values.

IV. Discussion on Multiple-Precision Arithmetic

The low-frequency breakdown, hence the requirement for a higher machine precision, occurs during the matrix vector multiplication. As a result, multiple-precision MLFMA implementations require the modification of the far-zone interactions. Moreover, we note that the required machine precision is strictly dependent on the translation distance (see Tables I and II); therefore, MPA should be hierarchically implemented across each translation level of MLFMA (i.e., $N-2$ different precisions for N levels). More specifically, setup, aggregation, translation, and disaggregation operations for each level must all be performed in the corresponding machine precision found by using the proposed method.

We illustrate the computational overhead introduced by the MPA for $\epsilon_d = 10^{-5}$ in Fig. 7, where we plotted the CPU-times and allocated memories for a one-box-buffer scheme. The simulations were performed on a workstation with 24-core Xeon E5-2650 processor. To have a fair comparison with the standard double precision, we computed the diagonal form of Green's functions using the same truncation numbers obtained from Table I for both double precision and MPA. An expected observation from Fig. 7 is that

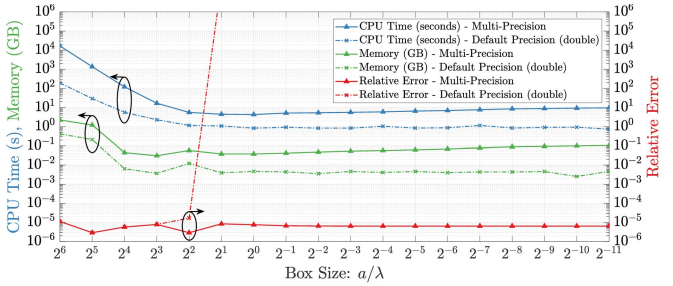

Fig. 7. Comparison of CPU-times, allocated memory, and achieved relative errors for varying box sizes when the truncation numbers in Table I is used for $\epsilon_d = 10^{-5}$ with double precision and MPA (averaged over 10 runs).

the double precision only works for box sizes larger than 4λ, which is in agreement with Table I. Another important observation is that there is a relatively constant overhead introduced by the MPA toolbox even for double or lower precisions. Moreover, the CPU and RAM requirements increase as the box size increases, since more and more terms need to be included in the summations and integrations in (2). For low frequencies or small boxes, the CPU and RAM requirements are relatively constant and only increases slightly for increasing machine precision (e.g., in Fig. 7, the machine precision increases from 13 to 273). We note that the commercial toolbox implements the MPA framework at the software level, while a hardware implementation would be more efficient and would introduce less overhead.

V. Conclusion

In this communication, a novel error control scheme for MLFMA that is valid at all frequencies and arbitrary desired error thresholds is introduced and demonstrated. The previous studies on the error control are limited to electrically large translation distances, relatively large error thresholds, and fixed machine precisions. The proposed scheme can be used to obtain the optimum truncation numbers and the machine precisions for any translation distance, given an arbitrary desired error threshold. Given the available machine precision and the translation distances, the proposed scheme can also be used to estimate the achievable error levels. Moreover, an MPA implementation of MLFMA with the proposed error control scheme can elegantly mitigate the well-known low-frequency breakdown problem while requiring no change in the underlying formulation.

Currently, MPA operations can be implemented with open-source or commercial libraries with small changes to the standard MLFMA codes. However, software implementations of arbitrary precision arithmetic introduce a constant but manageable overhead in terms of processing time and memory, which can be addressed with a low-level (i.e., hardware) implementation for increased computational efficiency.

References

[1] B. A. Cipra, "The best of the 20th century: Editors name top 10 algorithms," *SIAM News*, vol. 33, no. 4, pp. 1–22, May 2000.
[2] R. Coifman, V. Rokhlin, and S. Wandzura, "The fast multipole method for the wave equation: A pedestrian prescription," *IEEE Antennas Propag. Mag.*, vol. 35, no. 3, pp. 7–12, Jun. 1993.
[3] J. M. Song, C.-C. Lu, and W. C. Chew, "Multilevel fast multipole algorithm for electromagnetic scattering by large complex objects," *IEEE Trans. Antennas Propag.*, vol. 45, no. 10, pp. 1488–1493, Oct. 1997.
[4] L. Greengard, J. Huang, V. Rokhlin, and S. Wandzura, "Accelerating fast multipole methods for the Helmholtz equation at low frequencies," *IEEE Comput. Sci. Eng.*, vol. 5, no. 3, pp. 32–38, Jul. 1998.
[5] S. Koc, J. Song, and W. C. Chew, "Error analysis for the numerical evaluation of the diagonal forms of the scalar spherical addition theorem," *SIAM J. Numer. Anal.*, vol. 36, no. 3, pp. 906–921, Jan. 1999.

[6] J. M. Song and W. C. Chew, "Error analysis for the truncation of multipole expansion of vector Green's functions," *IEEE Microw. Wireless Compon. Lett.*, vol. 11, no. 7, pp. 311–313, Jul. 2001.

[7] W. C. Chew, J.-M. Jin, E. Michielssen, and J. M. Song, *Fast and Efficient Algorithms in Computational Electromagnetics*. Boston, MA, USA: Artech House, 2001.

[8] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. New York, NY, USA: Dover, 1964.

[9] M. L. Hastriter, S. Ohnuki, and W. C. Chew, "Error control of the translation operator in 3D MLFMA," *Microw. Opt. Technol. Lett.*, vol. 37, no. 3, pp. 184–188, 2003.

[10] J.-S. Zhao and W. C. Chew, "Three-dimensional multilevel fast multipole algorithm from static to electrodynamic," *Microw. Opt. Technol. Lett.*, vol. 26, no. 1, pp. 43–48, Jul. 2000.

[11] J.-S. Zhao and W. C. Chew, "Applying matrix rotation to the three-dimensional low-frequency multilevel fast multipole algorithm," *Microw. Opt. Technol. Lett.*, vol. 26, no. 2, pp. 105–110, Jul. 2000.

[12] J.-S. Zhao and W. C. Chew, "Applying LF-MLFMA to solve complex PEC structures," *Microw. Opt. Technol. Lett.*, vol. 28, no. 3, pp. 155–160, Feb. 2001.

[13] Y.-H. Chu and W. C. Chew, "A multilevel fast multipole algorithm for electrically small composite structures," *Microw. Opt. Technol. Lett.*, vol. 43, no. 3, pp. 202–207, Nov. 2004.

[14] Ö. Ergül and L. Gürel, "Efficient solutions of metamaterial problems using a low-frequency multilevel fast multipole algorithm," *Prog. Electromagn. Res.*, vol. 108, pp. 81–99, 2010.

[15] V. Melapudi, B. Shanker, S. Seal, and S. Aluru, "A scalable parallel wideband MLFMA for efficient electromagnetic simulations on large scale clusters," *IEEE Trans. Antennas Propag.*, vol. 59, no. 7, pp. 2565–2577, Jul. 2011.

[16] L. J. Jiang and W. C. Chew, "Low-frequency fast inhomogeneous plane-wave algorithm (LF-FIPWA)," *Microw. Opt. Technol. Lett.*, vol. 40, no. 2, pp. 117–122, Jan. 2004.

[17] I. Bogaert, J. Peeters, and F. Olyslager, "A nondirective plane wave MLFMA stable at low frequencies," *IEEE Trans. Antennas Propag.*, vol. 56, no. 12, pp. 3752–3767, Dec. 2008.

[18] I. Bogaert and F. Olyslager, "A low frequency stable plane wave addition theorem," *J. Comput. Phys.*, vol. 228, no. 4, pp. 1000–1016, Mar. 2009.

[19] Ö. Ergül and B. Karaosmanoğlu, "Low-frequency fast multipole method based on multiple-precision arithmetic," *IEEE Antennas Wireless Propag. Lett.*, vol. 13, pp. 975–978, 2014.

[20] Ö. Ergül and L. Gürel, *The Multilevel Fast Multipole Algorithm for Solving Large-Scale Computational Electromagnetics Problems*. Hoboken, NJ, USA: Wiley, 2014.

[21] *Multiprecision Computing Toolbox for MATLAB*. Accessed: Aug. 11, 2017. [Online]. Available: http://www.advanpix.com