

論文 / 著書情報
Article / Book Information

Title	Petascale turbulence simulation using a highly parallel fast multipole method
Author	Rio Yokota, Lorena Barba, Tetsu Narumi, Kenji Yasuoka
Journal/Book name	Computer Physics Communications, Vol. 184, No. 3, pp. 445–455
Issue date	2012, 9
URL	http://www.journals.elsevier.com/computer-physics-communications
DOI	http://dx.doi.org/10.1016/j.cpc.2012.09.011
Note	このファイルは著者（最終）版です。 This file is author (final) version.

Petascale turbulence simulation using a highly parallel fast multipole method on GPUs

Rio Yokota^a, L. A. Barba^{a,*}, Tetsu Narumi^b, Kenji Yasuoka^c

^aDept. of Mechanical Engineering, Boston University, Boston MA 02215

^bDept. of Computer Science, University of Electro-Communications, Tokyo, Japan

^cDept. of Mechanical Engineering, Keio University, Yokohama, Japan

Abstract

This paper reports large-scale direct numerical simulations of homogeneous-isotropic fluid turbulence, achieving sustained performance of 1.08 petaflop/s on GPU hardware using single precision. The simulations use a vortex particle method to solve the Navier-Stokes equations, with a highly parallel fast multipole method (FMM) as numerical engine, and match the current record in mesh size for this application, a cube of 4096^3 computational points solved with a spectral method. The standard numerical approach used in this field is the pseudo-spectral method, relying on the FFT algorithm as numerical engine. The particle-based simulations presented in this paper quantitatively match the kinetic energy spectrum obtained with a pseudo-spectral method, using a trusted code. In terms of parallel performance, weak scaling results show the FMM-based vortex method achieving 74% parallel efficiency on 4096 processes (one GPU per MPI process, 3 GPUs per node of the Tsubame-2.0 system). The FFT-based spectral method is able to achieve just 14% parallel efficiency on the same number of MPI processes (using only CPU cores), due to the all-to-all communication pattern of the FFT algorithm. The calculation time for one time step was 108 seconds for the vortex method and 154 seconds for the spectral method, under these conditions. Computing with 69 billion particles, this work exceeds by an order of magnitude the largest vortex-method calculations to date.

Keywords: isotropic turbulence, fast multipole method, integral equations, GPU

1. Introduction

In the history of using computer simulation as a research tool to study the physics of turbulence, the dominant approach has been to use spectral methods. Direct numerical simulation (DNS) was introduced as a means to check the validity of turbulence theories *directly* from the equations of fluid dynamics [7, p. 361]. The idea that important features of turbulence are *universal* encouraged researchers to study the simplest of geometries, a periodic cubic volume of homogeneous, isotropic turbulent fluid. In this case, the simplicity and efficiency of a Fourier-spectral method cannot be matched. The largest direct numerical simulation of isotropic turbulence was conducted by Ishihara et al. [14, 22] using 4096^3 grid points at a maximum Taylor-microscale Reynolds number $R_\lambda = 1200$. This record-breaking computation was done on *Earth Simulator*—a large vector machine with crossbar switch interconnect that can efficiently perform large-scale FFT. The successive generations of supercomputers have not been so FFT-friendly, and this record has not been surpassed even though the peak performance of supercomputers has increased nearly 50-fold since then. The record was matched for the first time in the U.S. by Donzis et al. [9], running on 16 thousand CPU cores of the Ranger super-

computer in Texas, a Linux-cluster supercomputer with 16-core nodes.

Future high-performance computing systems will have ever more nodes, and ever more cores per node, but will probably not be equipped with the bandwidth required by many popular algorithms to transfer the necessary data at the optimal rate. This situation is detrimental to parallel scalability. Therefore, it is becoming increasingly important to consider alternative algorithms that may achieve better sustained performance on these extremely parallel machines of the future.

In most standard methods of incompressible CFD, the greatest fraction of the calculation runtime is spent solving a Poisson equation. Equations of this type can be efficiently solved by means of an FFT-based algorithm, a sparse linear solver, or a fast multipole method (FMM) [11]. For the sake of our argument, we will not differentiate between FFT-based Poisson solvers and pseudo-spectral methods because they both rely on FFT. The fast multipole method has not gained popularity due to the fact that it is substantially slower—depending on implementations, at least an order of magnitude slower—than FFT and multigrid solvers when compared using a small CPU cluster. We aim to show with our ongoing research that the relative performance of FMM improves as one scales to large numbers of processes using GPU acceleration.

The highly scalable nature of the FMM algorithm, among other features, makes it a top contender in the algorithmic toolbox for exascale systems. One point of evidence for this argument is given by the Gordon Bell Prize winner of 2010, which

*Correspondence to: 110 Cummington St, Boston MA 02215, (617) 353-3883, labarba@bu.edu

Email addresses: yokota@bu.edu (Rio Yokota), labarba@bu.edu (L. A. Barba), narumi@cs.uec.ac.jp (Tetsu Narumi), yasuoka@mech.keio.ac.jp (Kenji Yasuoka)

achieved 0.7 petaflop/s with an FMM algorithm on 200k cores of the Jaguar supercomputer at Oak Ridge National Laboratory [17]. In the previous year, FMM also figured prominently at the Supercomputing Conference, with a paper among the finalists for the Best Paper award [16] and the Gordon Bell prize in the price/performance category going to work with hierarchical N -body methods on GPU architecture [12].

The FMM algorithm is well adapted to the architectural features of GPUS, which is an important consideration given that GPUS are likely to be a dominant player as we move towards exascale. The work presented in 2009—winner in the price/performance category in great measure thanks to the ingenious and resourceful system design using gaming hardware—reported (at the time of the conference) 80 teraflop/s [12]. That work progressed to an Honorable Mention in the 2010 list of awardees, with a performance of 104 teraflop/s [13].

At the level of the present work, where we present a 1.08 petaflop/s (single precision) calculation of homogeneous isotropic turbulence, the FMM moves firmly into the arena of *petascale* GPU computing. The significance of this advance is that we are now in a range where the FMM algorithm shows its true capability. The excellent scalability of FMM using over 4000 GPUS is an advantage over the dominant FFT-based algorithms. Showcasing the FMM in the simulation of homogeneous isotropic turbulence is especially fitting, given that a years-old record there remains unchallenged. We not only match the grid size of the world record, but also demonstrate that using the FMM as the underlying algorithm enables isotropic turbulence simulations that scale to thousands of GPUS. Given the severe bottleneck imposed by the all-to-all communication pattern of the FFT algorithm, this is not possible with pseudo-spectral methods in current hardware.

2. Integration of FMM into turbulence simulations

The FMM is used here as the numerical engine in a vortex particle method, which is an approach to solve the Navier-Stokes equations using a the vorticity formulation of momentum conservation and a particle discretization of vorticity. This is not a standard approach for the simulation of turbulence and the vortex method is not yet trusted for this application. For this reason, we have made efforts to document the validation of our vortex-method code and, in a separate publication [28], compared with a trusted spectral-method code. We looked at various turbulence statistics, including higher-order velocity derivatives, and performed a parameter study for the relevant numerical parameters of the vortex method. That work provides evidence that the vortex method is an adequate tool for direct numerical simulation of fluid turbulence, while in the present work we focus on the performance aspects. For completeness, this sections gives a brief overview of the numerical methods.

2.1. Vortex method

The vortex method [6] is a particle-based approach for fluid dynamics simulations. The particle discretization results in the

continuum physics being solved as an N -body problem. Therefore, the hierarchical N -body methods that extract the full potential of GPUS can be used for the simulation of turbulence. Unlike other particle-based solvers for fluid dynamics, e.g., smoothed particle hydrodynamics [10], the vortex method is especially well suited for computing turbulent flows, because the vortex interactions seen in turbulent flows are precisely what it calculates with the vortex particles.

In the vortex method, the Navier-Stokes equation is solved in the velocity-vorticity formulation, using a particle discretization of the vorticity field. The velocity is calculated using the following equation, representing the Biot-Savart law of fluid dynamics:

$$\mathbf{u}_i = \sum_{j=1}^N \gamma_j \times \nabla G g_\sigma. \quad (1)$$

Here, γ_j is the strength of vortex particles, $G = 1/4\pi r_{ij}$ is the Green's function for the Laplace equation and

$$g_\sigma = \text{erf} \left(\sqrt{\frac{r_{ij}^2}{2\sigma_j^2}} \right) - \sqrt{\frac{4}{\pi}} \sqrt{\frac{r_{ij}^2}{2\sigma_j^2}} \exp \left(-\frac{r_{ij}^2}{2\sigma_j^2} \right) \quad (2)$$

is the cutoff function, with r the distance between the interacting particles, and σ the standard deviation of the Gaussian function. The Navier-Stokes system is solved by a simultaneous update of the particle positions to account for convection, of the particle strengths to account for vortex stretching, and of the particle width to account for diffusion. The equation used to calculate the stretching term, $\omega \cdot \nabla \mathbf{u}$, is:

$$\frac{d\gamma_i}{dt} = \sum_{j=1}^n \nabla(\gamma_j \times \nabla G g_\sigma) \cdot \gamma_i, \quad (3)$$

which was obtained by substituting the Biot-Savart equation (1) for \mathbf{u} , and using the discrete form of the vorticity. Finally, the diffusion update is calculated according to

$$\frac{d\sigma^2}{dt} = 2\nu. \quad (4)$$

We perform a radial basis function interpolation for reinitialized Gaussian distributions to ensure the convergence of the diffusion calculation [1]. Equations (1) and (3) are N -body interactions, and are evaluated using the FMM. We use a highly parallel FMM library for GPUS developed in our group, called *exaFMM*, which is available under an open-source license and is described further below.

2.2. Spectral method

For the purpose of comparing with a spectral method, we used a code for homogeneous isotropic turbulence developed and used at the Center for Turbulence Research of Stanford University [5]. The code is called *hit3d* and is available freely for download.¹ It uses a spectral Galerkin method in primitive-variable formulation with pseudo-spectral methods to compute

¹In Google Code at <http://code.google.com/p/hit3d/>

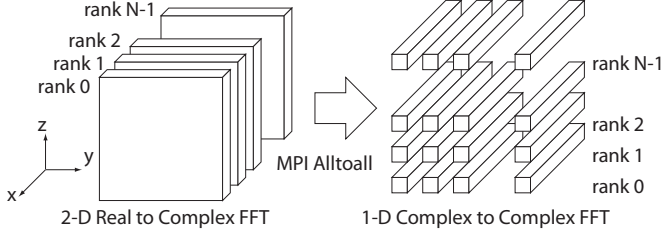


Figure 1: Two stage parallel FFT in spectral method calculation

the convolution sums. For the FFT, it relies on the `fftw` library and it provides parallel capability for CPU clusters using MPI. Parallelization of the FFT is accomplished by a domain-decomposition approach, illustrated by the schematic of Figure 1. Domain decomposition is applied in two spatial directions, resulting in first “slabs” then “pencil” sub-domains that are assigned to each MPI process.

The initial conditions for our runs were generated by `hit3d`. The vortex method used the same initial condition by first calculating the vorticity field in physical space, and then using radial basis function interpolation to obtain the vortex strengths. This is different from our other publication, where we studied the accuracy of turbulence simulations using the FMM-based vortex method on GPUs, looking at high-order turbulence statistics [28]. For that case, the initial condition provided by `hit3d`, which has a fully developed energy spectrum, was not suitable for our validation exercise looking at the time evolution of the velocity derivative skewness and flatness. For this reason, we constructed initial conditions in Fourier space as a solenoidal isotropic velocity field with random phases and a prescribed energy spectrum. This initial velocity field had a Gaussian distribution and satisfied the incompressibility condition.

3. Parallel fast multipole method on GPUs

It is common that algorithms with low complexity (sparse linear algebra, FFT) have low arithmetic intensity, while algorithms with high arithmetic intensity (dense linear algebra) tend to have high complexity. The FMM possesses a rare combination of $O(N)$ complexity and an arithmetic intensity that is even higher than `DGEMM` [23]. Although this may seem like a great combination, it also implies that there is a large constant in front of the $O(N)$ scaling, which results in a larger time-to-solution compared to other $O(N)$ or $O(N \log N)$ methods like multigrid methods and FFT. However, as arithmetic operations become cheaper compared to data movement in terms of both cost and energy, the large asymptotic constant of the *arithmetic* complexity becomes less of a concern.

3.1. Brief overview of the fast multipole method

The calculation of velocity by means of the Biot-Savart equation (1) and the computation of the stretching term (3) both result in an N -body problem, which has a complexity of $O(N^2)$ for N particles, if solved directly. The fast multipole method [11] reduces the complexity to $O(N)$ by clustering source and

target particles, and using series expansions that are valid far (multipole expansion) or near (local expansion) a point. The far/near relationships between points in the domain and the interactions between clusters are determined by means of a tree structure. The domain is hierarchically divided, and different sub-domains are associated with branches in the tree, a graphical representation of which is shown in Figure 2. The algorithm proceeds as follows. First, the strengths of the particles (e.g., charges or masses) are transformed to multipole expansions at the leaf cells of the tree (known as the particle-to-multipole, or $P2M$ kernel). Then, the multipole expansions of the smaller cells are translated to the center of larger cells in the tree hierarchy recursively and added (multipole-to-multipole, or $M2M$ kernel). Subsequently, the multipole expansions are transformed into local expansions for all pairs of well-separated cells (multipole-to-local, or $M2L$ kernel), and then to local expansions at the center of smaller cells recursively (local-to-local, or $L2L$ kernel). Finally, the local expansions at leaf cells are used to compute the effect of the far field on each target particle. Since $M2L$ operations can only be performed for well-separated cells, the direct neighbors at the finest level of the tree interact directly via the original equation (particle-to-particle, or $P2P$ kernel). The current implementation of the `exaFMM` code uses expansions in spherical harmonics of the Laplace Green’s function [4]. Details of the extension of Laplace kernels to Biot-Savart and stretching kernels and of the implementation on GPUs can be found in previous publications [25, 27].

3.2. Tree partitioning

When parallelizing hierarchical N -body algorithms, the fact that the particle distribution is dynamic makes it impossible to precompute the decomposition and to balance the work-load or communication *a priori*. Warren and Salmon [20] developed a parallel algorithm for decomposing the domain into recursive subdomains using the method of orthogonal recursive bisection (ORB). The resulting process can be thought of as a binary tree, which splits the domain into subdomains with equal number of particles at every bisection. Another popular technique for partitioning tree structures is to use Morton ordering [21], where bits representing the particle coordinates are interleaved to form a unique key that maps to each cell in the tree. Following the Morton index monotonically will take the form of a space-filling curve in the shape of a “Z”. Partitioning the list of Morton indices equally assures that each partition will contain an equal number of cells, regardless of the particle distribution.

For an adaptive tree, a naïve implementation of the Morton ordering could result in large communication costs if the “Z” is split in the wrong place, as illustrated in Figure 3. Sundar et al. [19] proposed a bottom-up coarsening strategy that ensures a clean partition at the coarse level while using Morton ordering. On the other hand, an ORB always partitions the domain into a well balanced binary tree, since the number of particles is always equal on each side of the bisection. Therefore, ORB is advantageous from the point of view of load-balancing.

The main difference between using Morton ordering and using an ORB is the shape of the resulting subdomains and the corresponding tree structure that connects them, as shown in

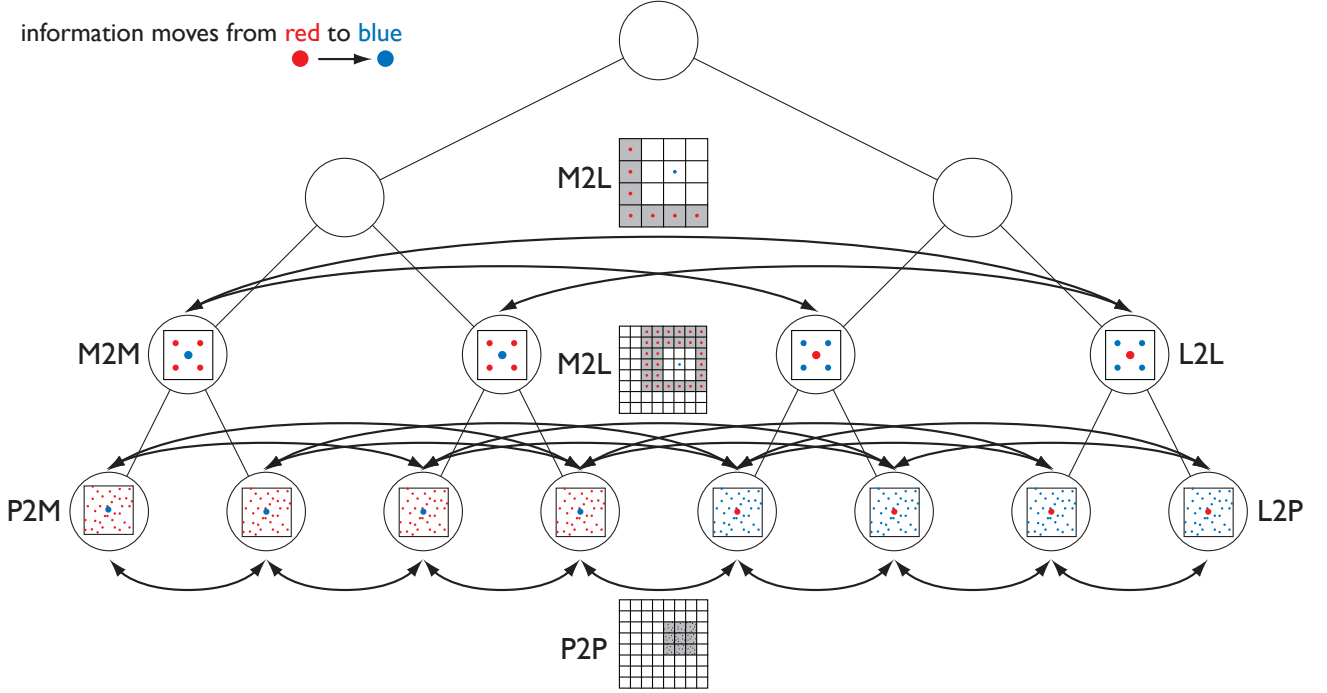


Figure 2: Illustration of the flow of an FMM calculation, with the naming convention used for the various computational kernels: particle-to-multipole or P2M; multipole-to-multipole or M2M; multipole-to-local or M2L; local-to-local or L2L; local-to-particle or L2P; particle-to-particle or P2P (all explained in the text).

Figure 3. Morton ordering is done on cubic octrees and the corresponding tree structure can become highly non-uniform depending on the particle distribution. Conversely, an ORB always creates a balanced tree structure but the sub-domain shapes are rectangular and non-uniform.

A difficulty when applying ORB within the framework of conventional FMM is that the construction of cell-cell interaction lists depends on the cells being cubic (and not rectangular). The dual tree traversal described in the following subsection allows the use of rectangular cells, while enabling cell-cell interactions in the FMM with minimum complications. By using ORB along with the dual tree traversal, we obtain a partitioning+traversal mechanism that allows perfect load balancing for highly non-uniform particle distributions, while retaining the $O(N)$ complexity of the FMM. Another advantage of this technique is that global indexing of cells is no longer necessary. Having a global index becomes an issue when solving problems of large size. The maximum depth of octrees that a 64-bit integer can handle is 21 levels ($2^{63} = 8^{21}$). For highly adaptive trees with small number of particles per cell, this limit will be reached quite easily, resulting in integer-overflow with a global index. Circumventing this problem by using multiple integers for index storage will require much more work when sorting, so this is not a desirable solution. The dual tree traversal allows the entire FMM to be performed without the use of global indices, and is an effective solution to this problem.

Our current partitioning scheme is an extension of the ORB, which allows multi-sections instead of bisections. Bisecting

the domain involves the calculation of the median of the particle distribution for a given direction, and doing this recursively in orthogonal directions (x, y, z, x, y, \dots) is what constitutes the “orthogonal recursive bisection”. Therefore the extension from bisection to multi-section can be achieved by simply providing a mechanism to search for something other than the median. We developed a parallel version of the “ n^{th} -element” algorithm. Finding the “ n^{th} -element” is much faster than any sorting algorithm, so our technique is much faster than any method that requires sorting. Always searching for the $N/2$ -th element will reduce to the original algorithm based on bisections, but searching for the $3N/7$ -th element will enable the domain to be split between 3 and 4 processes, for example. Therefore, our recursive multisection allows efficient partitioning when the number of processes is not a power of two.

3.3. Dual tree traversal

The dual tree traversal enables cell-cell interactions in the $O(N \log N)$ Barnes-Hut treecode [2] framework, thus turning it into a $O(N)$ algorithm [8]. We give a detailed explanation of the dual tree traversal in a previous publication that focused on hybrid treecode & FMM techniques [23]. It has several advantages compared to explicit construction of cell-cell interaction lists, the more common approach in the FMM framework. Firstly, it is the simplest and most general way to perform the task of finding all combinations of cell-cell interactions. It is simple in the sense that it does not contain multiple loops for finding “the parent’s neighbor’s children that are non-neighbors of the current

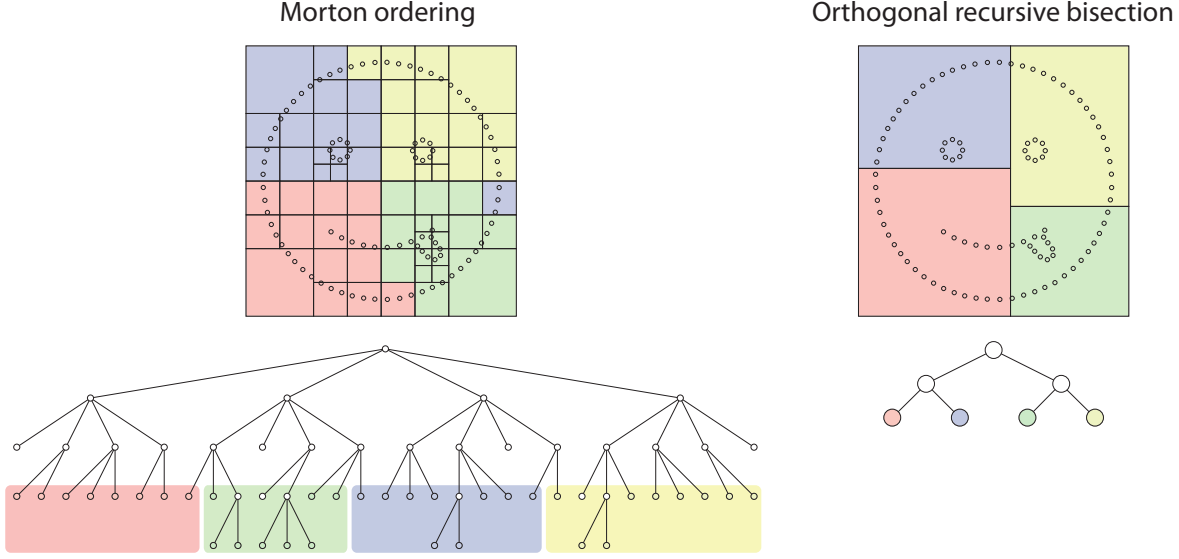


Figure 3: Example of partitioning with Morton ordering (left) or with orthogonal recursive bisection, ORB (right) for a non-uniform particle distribution. The ORB partition always gives a balanced tree.

cell,” as is the case with the explicit construction. It is general in the sense that the property “well-separated” can be defined flexibly, instead of the rigid definition of “non-neighboring cells” used traditionally in the FMM. It is common in treecodes to define the well-separated property more precisely and adaptively by introducing the concept of multipole acceptance criterion (MAC) [18].

We give a graphical description of the different types of MAC used in our method in Figure 4. The simplest MAC is the one defined for Barnes-Hut treecodes, where L_j is the size of the source cell and R is the distance from the target particle to the center of mass of the source cell; θ , called the opening angle, is the parameter that is used to flexibly control the definition of the well-separated property. In the FMM framework, a MAC is defined between two cells (rather than a cell and a particle), where L_i is the size of the target cell and R is the distance between the center of mass of the target cell and the center of mass of the source cell. A flexible definition of the well-separated property cannot be implemented easily in the traditional FMM framework, where the cell-cell interaction lists are constructed explicitly. This is because one must first provide a list of candidates to apply the MAC to, and increasing the neighbor search domain to $5^3 = 125$ neighbors instead of $3^3 = 27$ neighbors and applying the MAC on them is not an efficient solution. The dual tree traversal is a natural solution to this problem, since the list of candidates for the cell-cell interaction is inherited directly from the parent. This “inheritance of cell-cell interaction candidates” is lacking from conventional FMM, and can be provided by the dual tree traversal while simplifying the implementation at the same time. Furthermore, since the dual tree traversal naturally allows interaction between cells at different levels of the tree, it automatically finds the pair of cells that appear in U, V, W, X -lists for adaptive trees [16], but with much greater flexibility (coming from the MAC) and no overhead of generating the lists

(since it doesn’t generate any).

Another advantage of the dual tree traversal is that it enables the use of non-cubic cells. As we have explained in section 3.2, this allows using ORB partitioning in the FMM framework, which has superior load balancing properties compared to Morton ordering and removes the dependence on global indexing. As mentioned above, global Morton indices are a problem when using millions of cores and the tree depth exceeds 21 levels causing the Morton index to overflow from the 64-bit integer. The combination of the dual tree traversal and ORB is an elegant solution to this problem as well.

We describe the dual tree traversal in Algorithm 1, which calls an internal routine for the interaction of a pair of cells, given in Algorithm 2. First, a pair of cells is pushed to a stack. It is most convenient to start from the root cell although there is no possibility that two root cells will interact. For every step in the while-loop, a pair of cells is popped from the stack and the larger cell is subdivided. Then, Algorithm 2 is called to perform either particle-particle (P2P) or multipole-local (M2L) interactions. If the cells in the pair are too close and either of the cells has children, the pair is pushed to the stack and will be handled later. A more detailed explanation is given in a previous publication focusing on hybrid treecode & FMM techniques [23].

3.4. Local essential tree

The partitioning techniques discussed in section 3.2 will assign a local portion of the global tree to each process. However, the FMM evaluation requires information from all parts of the global tree, and the multipole expansions on remote processes must be communicated. Unlike standard domain-decomposition methods, where only a thin halo needs to be communicated, the FMM requires a global halo. Fortunately, this global halo becomes exponentially coarser as the distance

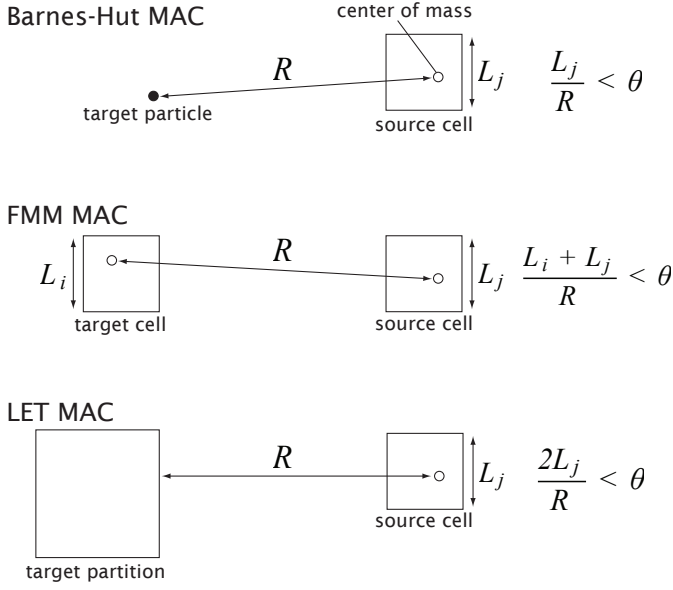


Figure 4: Definition of three types of multipole acceptance criterion (MAC). The Barnes-Hut MAC is the simplest version from the original treecode. The FMM MAC is defined between two cells rather than between a cell and particle. The LET MAC is used for the construction of the local essential tree (LET), and is defined between a cell and a partition on a remote process.

from the target cell increases, as shown in Figure 6. Therefore, the data required to be communicated from far partitions is very small (but never zero). This results in a non-homogeneous all-to-all-type communication of subsets of the global tree.

Once all the subsets of the global tree are communicated between the processes, one can locally reconstruct a tree structure that contains all the information required for the evaluation in the local process. This reconstructed tree is called the *local essential tree* (LET), and it is a (significantly smaller) subset of the global tree that contains all the information that is necessary to perform the evaluation. Salmon and Warren [20] introduced two key techniques in this area: one for finding which cell data to send, and the other for communicating the data efficiently.

The determination of which data to send is tricky, since each process cannot see what the adaptive tree structure looks like on remote processes. The solution proposed by Salmon and Warren [20] is to use a conservative estimate and communicate a larger portion of the tree than is exactly required. This conservative estimate is obtained by means of a special MAC described in Figure 4 as the LET MAC. There, the distance R is defined as the distance between the center of mass of the source cell and the edge of the partition on the remote process. A formula to calculate R can be given as the following expression, where we assume element-wise Boolean operations over array elements (giving zero for false, and 1 for true) and element-wise multiplication and summation:

$$R = |\Delta \mathbf{x}|$$

$$\Delta \mathbf{x} = (\mathbf{x} > \mathbf{x}_{max})(\mathbf{x} - \mathbf{x}_{max}) + (\mathbf{x} < \mathbf{x}_{min})(\mathbf{x} - \mathbf{x}_{min})$$

Algorithm 1 Evaluate()

```

A=B=rootcell
push pair(A,B) into a stack
while the stack is not empty do
  Pop stack to get a pair(A,B)
  if radius of A > radius of B then
    for all child cells "a" of cell A do
      Interact(a,B)
    end for
  else
    for all child cells "b" of cell B do
      Interact(A,b)
    end for
  end if
end while

```

Algorithm 2 Interact(cell A, cell B)

```

if there are very few particles in both cells, or they don't have
child cells then
  if particles weren't sent from remote processes then
    Evaluate multipole-local (M2L)
  else
    Evaluate particle-particle (P2P)
  end if
else if cells A and B are well separated then
  Evaluate multipole-local (M2L)
else
  Push pair(A,B) into a stack
end if

```

Here, \mathbf{x}_{min} and \mathbf{x}_{max} are the minimum and maximum coordinate values for all particles in the target partition, while \mathbf{x} is the center of mass of the source cell. This definition of the LET MAC corresponds to assuming the case where the target cell is located at the edge of the remote partition. We also assume that the target cell is of the same size as the source cell ($L_i = L_j$), so $L_i + L_j$ becomes $2L_j$. These assumptions generally hold quite well and the required part of the tree is sent to the remote process most of the time. However, we must ensure that the FMM code still works for extreme cases where necessary information fails to be sent. With this end in view, we have added a conditional statement in the interaction calculation (Algorithm 2) as a further precaution for anomalous cases. This way, the traversal will perform the M2L translation with the smallest cell that is available. This cell may be too large to satisfy the FMM MAC so there will be a small penalty on the accuracy, but since the occurrence of such a case is so rare, it does not affect the overall result.

A schematic showing how the partitioning and communication techniques are used in the vortex method is given in Figure 5. First, the domain is partitioned using the recursive multisec-tion described in section 3.2. Then the FMM kernels for the local tree are evaluated while the LET data is being communicated in a separate OpenMP section. After the LET data is communicated, the FMM kernels are evaluated again for the remaining parts of

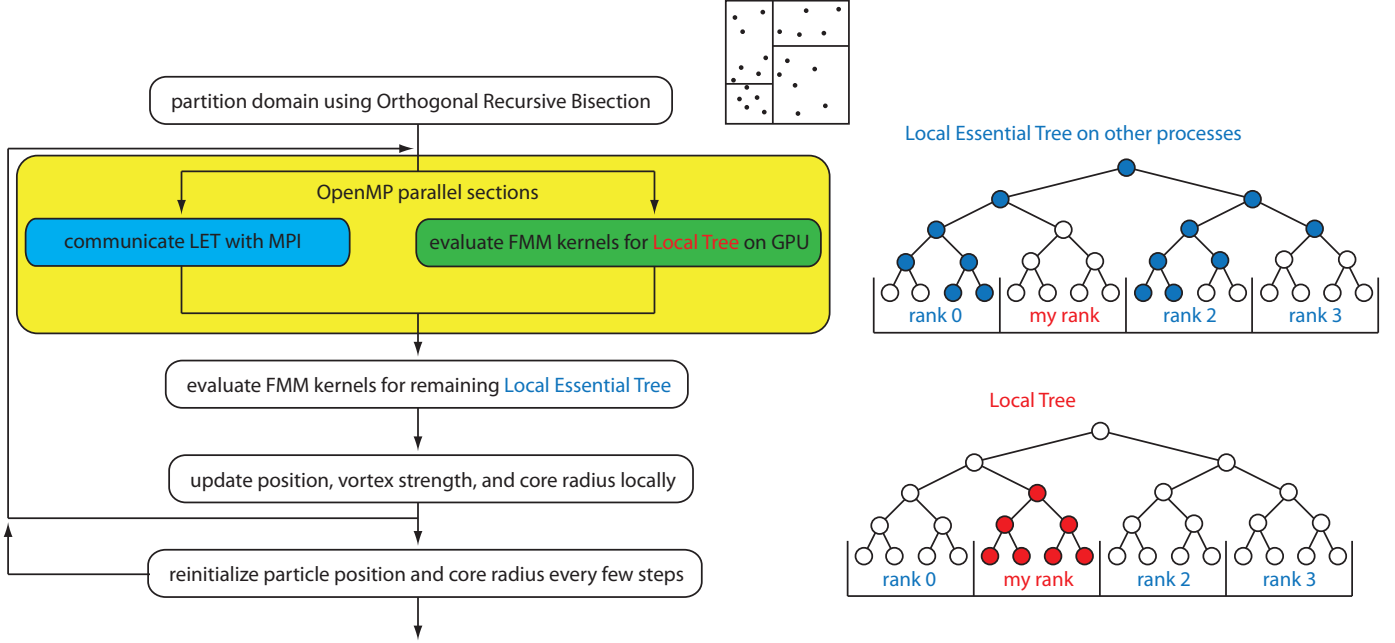


Figure 5: Flowchart explaining the use of partitioning and communication strategies of the FMM in a vortex method application code (left) and schematic of the local essential tree (right).

the LET. Subsequently, the position, vortex strength γ , and core radius σ of the vortex particles are updated locally. This information is communicated in the next stage when the LET is exchanged. In addition, the Lagrangian vortex method needs to reinitialize the particle positions to maintain sufficient overlap of the smooth particles. For this procedure, we can reuse the same tree structure since the particles are reinitialized to the same position every time. Therefore, the partitioning is performed only once at the beginning of this simulation.

3.5. Periodic FMM

The FMM was originally devised to solve potential problems with free-field boundary conditions. The method can be extended to handle periodic boundary conditions by placing periodic images around the original domain and using multipole expansions to approximate their influence [15], as illustrated in Figure 6. When a sufficient number of periodic images are placed, the error caused by using a finite number of periodic images becomes smaller than the approximation error of the FMM itself [15]. This approach to extend the FMM to periodic domains adds negligible computational overhead to the original FMM, for two reasons. First, distant domains are clustered into larger and larger cells, so the extra cost of considering a layer of periodic images is constant, while the number of images accounted for grows exponentially. The second reason is that only the sources need to be duplicated and the target points exist only in the original domain. Since the work load for considering the periodicity is independent of the number of particles, it becomes negligible as the problem size increases. An earlier study showed that periodic boundary conditions add approximately 3% to the calculation time for a million particles [26].

3.6. GPU kernels

The FMM consists of six different computational kernels, as illustrated on Figure 2. In the exaFMM code, all of these kernels are evaluated on GPU devices using CUDA. Out of the six kernels, a great majority of the runtime is spent executing P2P and M2L. We use a batch evaluation for these two kernels, since there is no data dependency between them: they are evaluated in one batch after the tree traversal is completed. This batch evaluation can be broken into multiple calls to the GPU device, depending on its storage capacity and the data size. With this approach, we are able to handle problem sizes of up to 100 million particles on a single GPU, if the memory on the host machine is large enough [29].

As an example of the usage of thread blocks in the GPU execution model, we show in Figure 7 an illustration of the M2L kernel on GPUs. Each coefficient of the multipole/local expansion is mapped to a thread on the GPU and each target cell is mapped to a thread block, while each source cell is loaded to shared memory and evaluated sequentially. All other kernels are mapped to the threads and thread blocks in a similar manner. More details regarding the GPU implementation of FMM kernels can be found in chapter 9 of the *GPU Gems Emerald Edition* book [27], with accompanying open-source codes².

4. Results: parallel simulations of isotropic turbulence

In this section, we present results from large-scale simulations of homogeneous isotropic turbulence using the FMM-based

²In Google Code at <http://code.google.com/p/gemsfmm/>

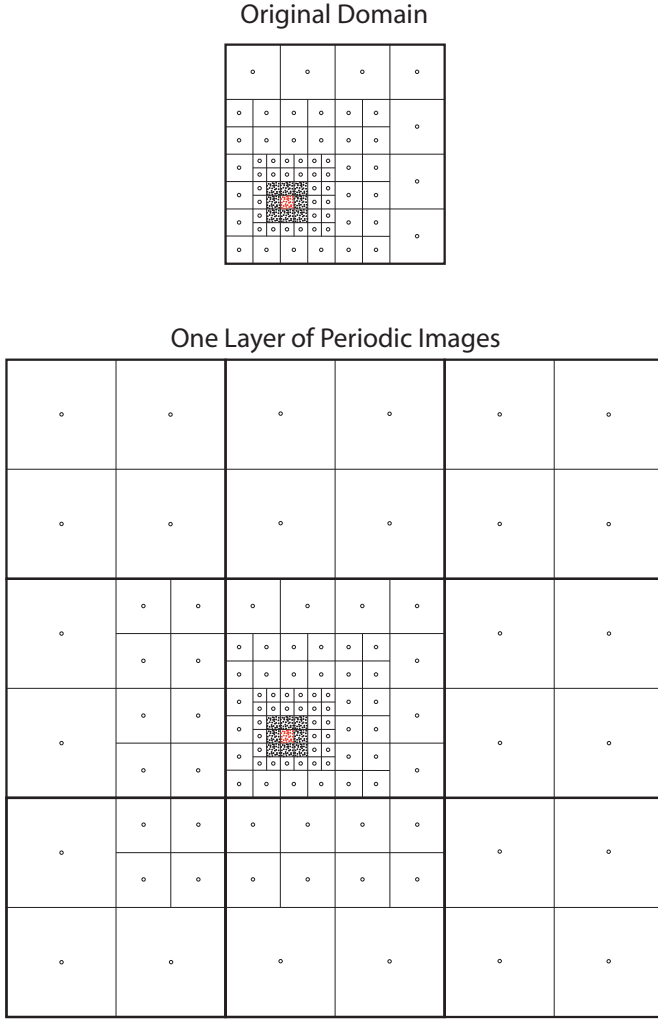


Figure 6: Using periodic image domains with the FMM interaction list, for a target cell illustrated by the red dots.

vortex particle method, up to a problem size of 4096^3 computational points. This is still the largest mesh-size for which DNS of turbulence has been published, even though this scale of simulation was achieved 10 years ago. As we discussed in the Introduction, one of the reasons for this is the difficulty in scaling the FFT algorithm beyond a few thousand processes. The current simulations with a vortex method were checked for correctness by comparing the kinetic energy spectrum with that obtained using a trusted spectral-method code (described in §2.2). The focus here is not on the physics, however, but on demonstrating large-scale simulation of turbulence using the FMM as a numerical engine and reporting on performance aspects using GPU hardware. The performance is described via the results of weak scaling tests using between 1 and 4096 processes with 256^3 (16.8 million) particles per process. Each process offloads to one GPU to speed-up the FMM, making it even more challenging to obtain good parallel efficiency (it is obviously harder to scale faster code). Despite this, a parallel efficiency of 74% was obtained for the FMM-based vortex method on weak scaling be-

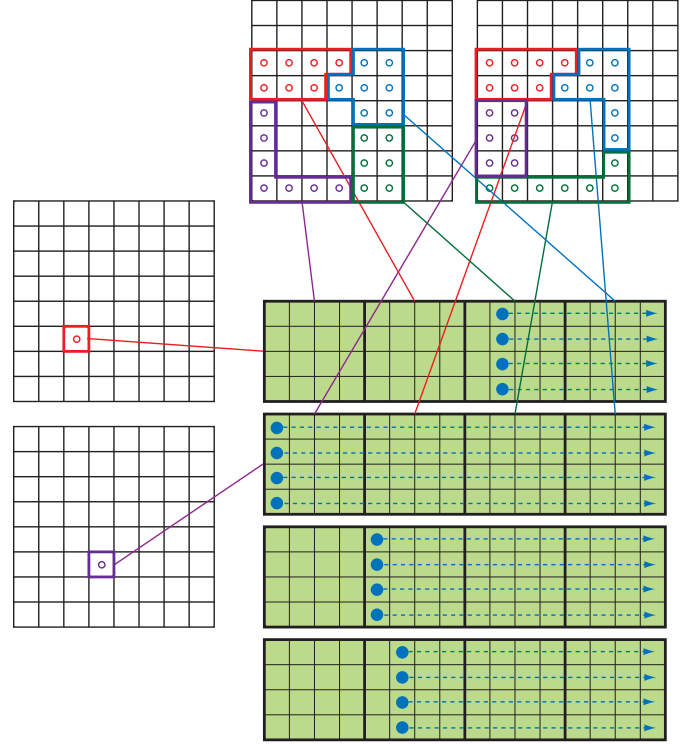


Figure 7: Execution of m2L kernel on GPUs, where each expansion coefficient is mapped to a thread and each target cell is mapped to a thread block. Figure appeared in the chapter by Yokota and Barba [27]—permission pending.

tween 1 and 4096 processes, with the full application code. The largest calculation used 69 billion particles, which as far as we know is the largest vortex-method calculation to date. Previous noteworthy results by other authors reported calculations with up to 6 billion particles [3], which we surpass here by an order of magnitude.

4.1. Hardware

The calculations reported here were run on the TSUBAME-2.0 system during Spring and Fall of 2011, thanks to guest access provided by the Grand Challenge Program of TSUBAME-2.0. This system has 1408 nodes, each equipped with two six-core Intel Xeon X5670 (formerly Westmere-EP) 2.93GHz processors, three NVIDIA M2050 GPUs, 54 GB of RAM (96 GB on 41 nodes), and 120 GB of local SSD storage (240 GB on 41 nodes). Computing nodes are interconnected with the InfiniBand device Grid Director 4700 developed by Voltaire Inc., with non-blocking and full bisectional bandwidth. Each node has 2×40 Gbps bandwidth, and the bisection bandwidth of the system is over 200 Tbps. The total number of M2050 GPUs in the system is 4224, and the peak performance of the entire system is 2.4 petaflop/s.

4.2. Test conditions

We set up simulations of decaying, homogeneous isotropic turbulence using an initial Taylor-scale Reynolds number of

$Re_\lambda \approx 500$. Given that we had guest access to the full TSUBAME-2.0 system for a very brief period of time (only a few hours) to produce the scalability results, we opted for a lower Reynolds number than previous turbulence simulations of this size using spectral methods. There is limited experience using vortex methods for DNS of turbulence, but previous work has suggested that higher resolutions are needed than when using the spectral method at the same Reynolds number. We thus decided to be conservative and ensure that we obtained usable results from these one-off runs.

The calculation domain is a box of size $[-\pi, \pi]^3$ with periodic boundary conditions in all directions, using 3^3 periodic images in each dimension in the periodic FMM (see Figure 6). To achieve the best accuracy from the FMM for the present application, the order of multipole expansions was set to $p = 14$; this may be a conservative value, but given our limited allocation in the TSUBAME-2.0 system, we did not have the luxury of tuning the simulations for this parameter.

The FMM kernels are run in single precision on the GPU which may raise some concerns. We are able to achieve double-precision accuracy using single-precision computations in the FMM kernels by means of two techniques. The multipole-expansion coefficients have exponentially smaller magnitude with increasing order of expansion; therefore, by adding them from higher- to lower-order terms, we can prevent small numbers being added to the large numbers. This preserves the significant digits in the final sum. The second technique consists of normalizing the expansion coefficients to reduce the dynamic range of the variables, allowing the use of single-precision variables to get a double-precision result. Thus, we are able to achieve sufficient accuracy to reproduce the turbulence statistics (as shown below) and obtain the same results that would be obtained using double precision (given that the FMM error is larger than 6 significant digits anyway).

4.3. Isosurface of the second invariant and kinetic energy spectrum of turbulence

The isosurface of the second invariant of the velocity gradient tensor is shown in Figure 9. This is a snapshot at the early stages of the simulation and we do not observe any large coherent structures. In order to take a closer look at the quantitative aspects of the vortex simulation, in Figure 8 we compare the kinetic energy spectrum with that of the spectral method, where T is the eddy turnover time. The energy spectrum of the vortex method is obtained by calculating the velocity field on a uniform lattice, which is induced by the Lagrangian vortex particles. The capability of exaFMM to calculate for different sources and targets enabled such operations. We have excellent quantitative agreement between the vortex method and spectral method and we conclude that our FMM-based particle method is capable of simulating turbulence of this scale correctly. We turn our attention to the performance of these calculation in the next section.

4.4. Weak scaling

We had two very short windows of time in which we were able to run weak scaling tests, one with half the system and

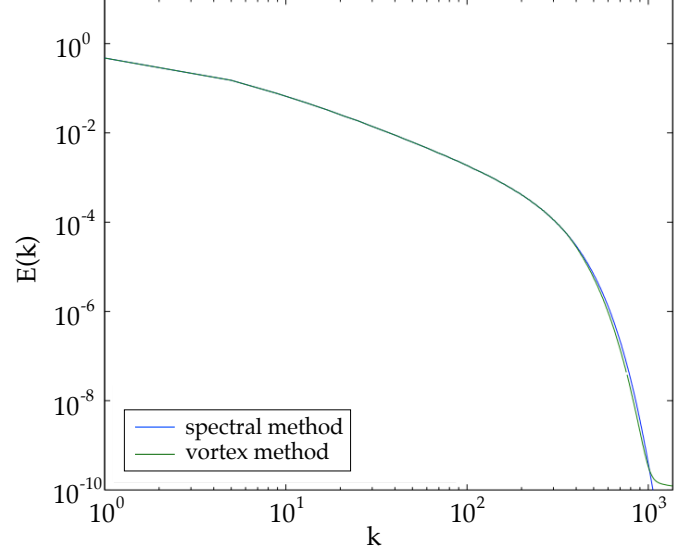


Figure 8: Kinetic energy spectrum at $t/T = 2$, obtained with the vortex method and spectral method. Notice that the vertical axis goes down to 10^{-10} , which is many times smaller than in similar plots presented by other authors.

the other with almost the full TSUBAME-2.0 system. The larger scaling tests used 256^3 particles per process, on 1 to 4096 processes and executing three MPI processes per node, each process assigned to a GPU card within the node. The results of the weak scaling test are shown in Figure 10 in the form of total runtime of the FMM, and timing breakdown for different phases in the computation. The label ‘Near-field evaluation’ corresponds to the P2P kernel evaluation illustrated in Figure 2, and the label ‘Far-field evaluation’ corresponds to the sum of all the other kernel evaluations, i.e., the far field. The ‘MPI communication’ is overlapped with the FMM evaluation (see Figure 5), so the plot shows only the amount of time that communication exceeds the local portion of the ‘near-field’ and ‘far-field’ evaluation. In this way, the total height of each bar correctly represents the total wall-clock time of each calculation. Note that particle updates in the vortex-method calculation take less than 0.01% in all runs and thus were invisible in the bar plots, so we’ve left this computing stage out of the labels.

The ‘GPU buffering’ label corresponds to the time it takes to form a queue of tasks and corresponding data buffer to be transferred to the GPU, which is a significant amount of time. We have found this buffering to be necessary in order to achieve high efficiency in the P2P and M2L evaluation and FMM evaluation on GPUs. Moreover, this part of the computation scales perfectly, and does not affect the scalability of the FMM. The parts that do affect the scalability are the tree construction and MPI communication. Actually, the tree construction also involves MPI communications for the partitioning, so the parallel efficiency in weak scaling is fully determined by MPI communications. Figure 10 shows that the current FMM is able to almost completely hide the communication time up to 512 GPUs.

It may be worth noting that the N -D-hypercube-type com-

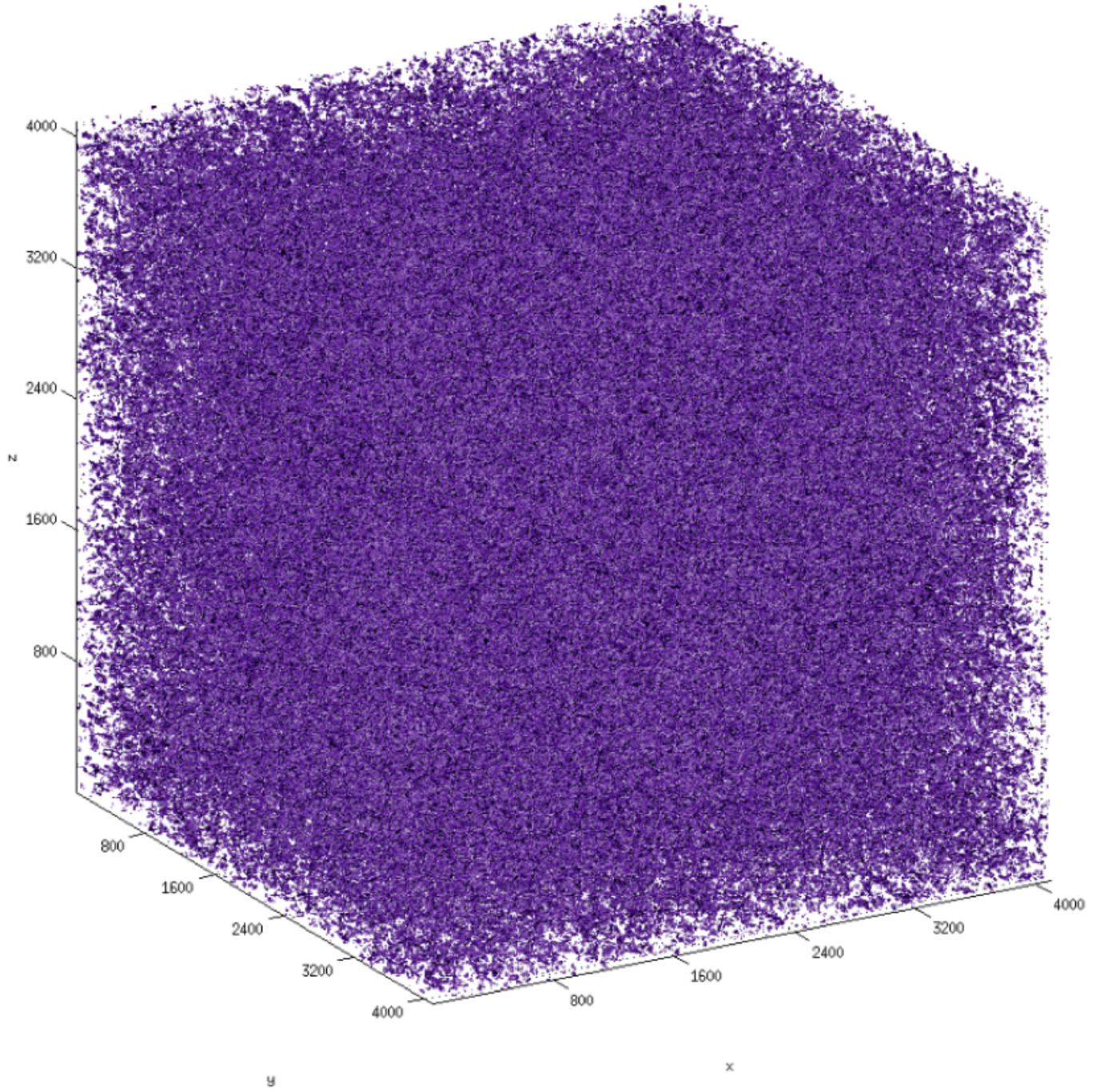


Figure 9: Plot of an isosurface of the second invariant of the velocity gradient tensor $II = -10$ for the 4096^3 mesh at $t/T = 2$.

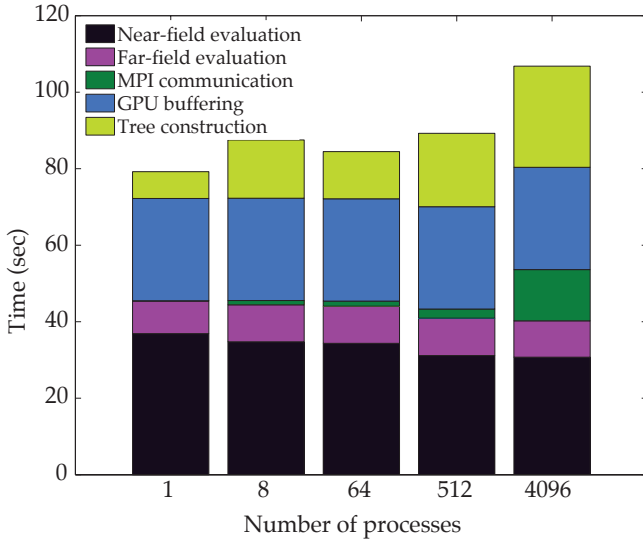


Figure 10: Runtime of weak scaling tests, using one GPU per process and three GPUS per node. All the bars would have the same height if the scaling was perfect. The bars also show a timing breakdown for important phases of the calculation; communication time is overlapped, and is almost completely hidden up to 512 processes.

munication of the LET [16] turned out to be slower on TSUBAME-2.0 than a simple call to `MPI_Alltoallv` for sending the entire LET at once. This is a consequence of the network topology of TSUBAME-2.0 (with a dual-QDR InfiniBand link to each node and non-blocking full-bisection fat-tree interconnect) and also of the relatively small number of MPI processes. The results shown in Figure 10 are those obtained with `MPI_Alltoallv` and not the hypercube-type communication.

We performed a corresponding weak scalability test for the spectral method, increasing the problem size from 256^3 on one process to 4096^3 on 4096 processes. We used three MPI processes per node to match the condition of the vortex method runs, but there is no GPU acceleration in this case. Matching the number of MPI processes per node should give both methods an equal advantage/handicap for the bandwidth per process. Note that using GPUs for the FFT within the spectral-method code is unlikely to provide any benefits, because performance improvements of `cufft` over `fftw` would be canceled out by data transfer between host and device³ and inter-node communications in parallel runs. Figure 11 shows the parallel efficiency obtained with the two methods, under these conditions.

The parallel efficiency of the FMM-based vortex method is 74% when going from one to 4096 GPUs, while the parallel efficiency of the spectral method is 14% when going from one to 4096 CPUs. The bottleneck of the spectral method is the all-to-all communication needed for transposing the slabs into pencils as shown in Figure 1. Even though this may not be the best implementation of a parallel FFT, the difference in the scalability between the spectral method and FMM-based vortex method is

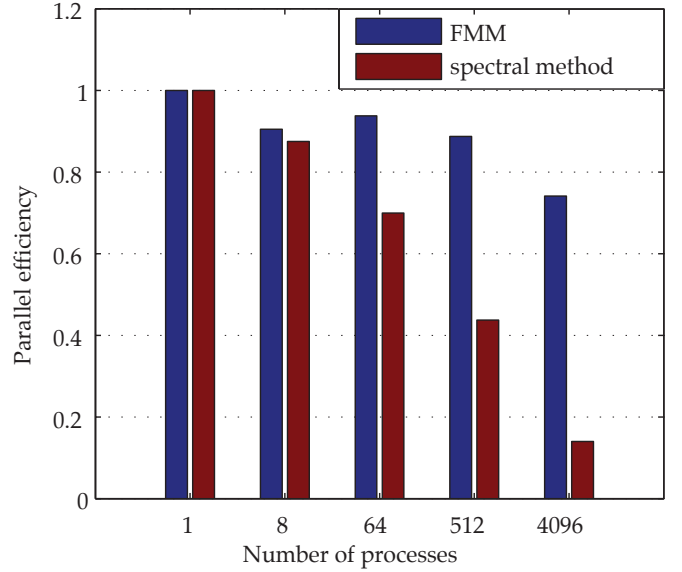


Figure 11: Weak scaling from 1 to 4096 processes of the parallel FMM-based fluid solver on GPUs, and of the pseudo-spectral method on CPUs. The parallel efficiency of the FMM-based solver is 74% at 4096 processes (one GPU per MPI process, 3 GPUS per node). The tests were ran on the TSUBAME-2.0 system in October 2011 using revision 191 of the exaFMM code. Figure, plotting script and dataset available online and usage licensed under CC-BY-3.0 [24].

considerable. The actual calculation time is in the same order of magnitude for both methods at 4096 processes: it was 108 seconds per time step for the vortex method and 154 seconds per time step for the spectral method. Therefore, the superior scalability of the FMM has merely closed the gap with FFT, being barely 1.44× faster at this scale. However, we anticipate that this trend will affect the choice of algorithm in the future, as we move to architectures with higher degree of parallelism.

The scaling test with half the TSUBAME-2.0 system was done several months before and with a different revision of the code, with many changes having been incorporated since then. We include the results here for completeness; see Figure 12. In this case, the number of particles per process is much smaller, at 4 million (compared to 16.8 million particles per process in the larger test) and we scale from 4 to 2048 processes. The parallel efficiency of the fmm-based vortex method is 72% when going from 4 to 2048 GPUs, while the parallel efficiency of the spectral method was 14%. When we compare the parallel efficiency with the 1 to 4096 GPUCASE, we see that the 4 to 2048 case is scaling relatively poorly. This is due to the number of particles per process being roughly 1/4 in the 4 to 2048 case, and also the improvement in the exaFMM code during the half year gap between the two runs. The run time per time step in this case was 27 seconds for the vortex method and 20 seconds for the spectral method to compute a domain with 2048^3 points. Note that if we read from the plot in Figure 2 of Donzis et al. [9], their spectral-method calculation on a 2048^3 grid using 2048 cores

³See http://www.sharcnet.ca/?merz/CUDA_benchFFT/

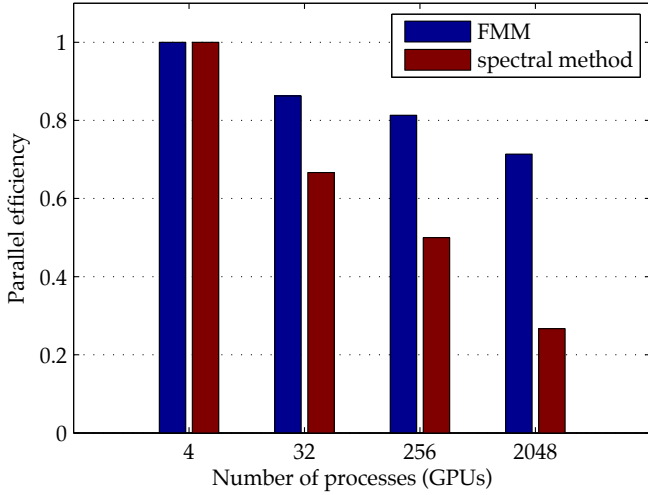


Figure 12: Weak scaling from 4 to 2048 processes of the parallel FMM-based fluid solver on GPUS, and of the pseudo-spectral method on CPUS, using 4 million particles per process. The parallel efficiency on 2048 processes is 72 % for the FMM-based solver, and 27 % for the FFT-based solver. These tests were run on the TSUBAME-2.0 system in April 2011, using revision 146 of the exaFMM code.

of the Ranger supercomputer takes them about 20 seconds per time step. Since this is the same timing we obtained with the hit3d code, we are satisfied that this code provides a credible reference point for the scalability of spectral DNS codes.

4.5. Sustained performance

The calculation in the FMM is mostly dominated by the floating point operations in the particle-particle interactions, while all other parts are a minor contribution in terms of flop/s (although not negligible in terms of runtime). We will thus consider in the estimate of sustained performance only the operations executed by the P2P kernels. Two separate equations are being calculated for the particle-particle interactions: the Biot-Savart equation (1) and the stretching equation (3). The number of floating point operations required by these two kernels is summarized in Table 1.

Table 1: Floating point operations per P2P interaction.

Operation	Biot-Savart	Stretching
+	19	25
-	14	18
*	32	56
/	1	1
sqrtf	1	1
rsqrtf	1	1
expf	2	2
Total	70	104

The approximate number of flop/s for one step of the vortex

method calculation of isotropic turbulence is obtained by the following equation.

$$\begin{aligned}
\text{flop/s} &= (\# \text{ processes}) \\
&\times (\text{target particles per process}) \\
&\times (\text{source cells per target}) \\
&\times (\text{source particle per cell}) \\
&\times (\text{flop/s per interaction}) \\
&/ (\text{wall clock time}) \\
&= 4096 \times (6.8 \times 10^7) \times 19 \times 512 \times 174 / 108 \\
&= 1.08 \times 10^{15} = \mathbf{1.08 \text{ petaflop/s}}
\end{aligned}$$

Thus, the current FMM-based vortex method achieved a sustained performance of 1.08 petaflop/s (single precision) on 4096 GPUS of TSUBAME-2.0.

4.6. Reproducibility and open-source policy

The authors of the exaFMM code have a consistent policy of making science codes available openly, in the interest of reproducibility. The entire code that was used to obtain the present results is available from <https://bitbucket.org/exafmm/exafmm>. The revision number used for the results presented in this paper is 191 for the large-scale tests up to 4096 GPUS. Documentation and links to other publications are found in the project homepage at <http://exafmm.org/>. Figure 11, its plotting script and datasets are available online and usage is licensed under CC-BY-3.0 [24].

We acknowledge the use of the hit3D pseudo-spectral DNS code for isotropic turbulence, and appreciate greatly their authors for their open-source policy; the code is available via Google code at <http://code.google.com/p/hit3d/>.

5. Conclusions

This work represents several milestones. Although the FMM algorithm has been taken to petascale before (notably, with the 2010 Gordon Bell prize winner), the present work represents the first time that this is done on GPU architecture. Also, to our knowledge, the present work is the largest direct numerical simulation with vortex methods to date, with 69 billion particles used in the cubic volume; this is an order of magnitude larger than the previously reported record. Yet another significant event is reaching a range where the highly scalable FMM starts showing advantage over FFT-based algorithms. With a 1.08 petaflop/s (single precision) calculation of isotropic turbulence in a 4096^3 box, using 4096 GPUS, we are within reach of a turning point. The combination of application, algorithm, and hardware used are also notable.

The real challenge in exascale computing will be the optimization of data movement. When we compare the data movement of FMM against other fast algorithms like multigrid and FFT, we see that the FMM has a potential advantage. Compared to FFT, both multigrid and FMM have an advantage in the asymptotic complexity of the global communication. The hierarchical nature of multigrid and FMM results in $O(\log P)$ communication complexity where P is the number of processes. On the

other hand, FFT requires two global-transpose communications between \sqrt{P} processes, and has communication complexity of $O(\sqrt{P})$. When P is in the order of millions, it seems unrealistic to expect that an affordable network can compensate for this large gap in the communication complexity. Although it is not the focus of the present article, we would like to briefly note that an advantage of FMM over multigrid is obtained from differences in the synchronization patterns. For example, increasing the desired accuracy in iterative solvers using multigrid will result in more iterations, hence more global synchronizations. Conversely, increasing the accuracy in FMM involves increasing the number of multipole expansions, which results in even higher arithmetic intensity in the inner kernels while the number of global synchronizations remains the same. As the amount of concurrency increases, bulk-synchronous execution/communication models are reaching their limit. Thus, FMM provides a new possibility to reduce the amount of communication and synchronization in these inherently “global” problems.

Finally, we would like to point out that the FMM can be used to solve the Poisson equation directly, or as a preconditioner for an iterative solver. Therefore, we are not concerned at this point about the fact that vortex methods may still be comparatively inefficient for the simulation of fluid turbulence, where spectral methods will continue to dominate in the foreseeable future. This does not detract from the conclusions about the efficiency of the FMM itself, which is the object of our study. In future work, we would like to demonstrate the efficiency of FMM by using it as a Poisson solver or preconditioner in the framework of more standard finite difference/volume/element methods. There, the comparison against FFT and multigrid methods should be of interest to a broader spectrum of the CFD community.

6. Acknowledgments

Computing time in the TSUBAME-2.0 system was made possible by the Grand Challenge Program of TSUBAME-2.0. The current work was partially supported by the Core Research for the Evolution Science and Technology (CREST) of the Japan Science and Technology Corporation (JST). LAB acknowledges funding from NSF grant OCI-0946441, ONR grant #N00014-11-1-0356 and NSF CAREER award OCI-1149784. LAB is also grateful for the support from NVIDIA Corp. via an Academic Partnership award (Aug. 2011).

References

- [1] L. A. Barba. *Vortex method for computing high-Reynolds number flows: Increased accuracy with a fully mesh-less formulation*. PhD thesis, California Institute of Technology, 2004.
- [2] J. Barnes and P. Hut. A hierarchical $O(N \log N)$ force-calculation algorithm. *Nature*, 324:446–449, December 1986.
- [3] P. Chatelain, A. Curioni, M. Bergdorf, D. Rossinelli, W. Andreoni, and P. Koumoutsakos. Billion vortex particle direct numerical simulations of aircraft wakes. *J. Comp. Phys.*, 197(13–16):1296–1304, 2008.
- [4] H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *J. Comput. Phys.*, 155:468–498, 1999.
- [5] S. G. Chumakov, J. Larsson, C. Schmitt, and H. Pitsch. Lag-modeling approach for dissipation terms in large eddy simulation. Annual research briefs, Center for Turbulence Research, 2009. <http://www.stanford.edu/group/ctr/ResBriefs/ARB09.html>.
- [6] G.-H. Cottet and P. Koumoutsakos. *Vortex Methods. Theory and Practice*. Cambridge University Press, 2000.
- [7] P.A. Davidson, Y. Kaneda, K. Moffatt, and K.R. Sreenivasan. *A voyage through turbulence*. Cambridge University Press, 2011.
- [8] W. Dehnen. A hierarchical $O(N)$ force calculation algorithm. *J. Comput. Phys.*, 179(1):27–42, 2002.
- [9] D. A. Donzis, P. K. Yeung, and D. Pekurovsky. Turbulence simulations on $O(N^4)$ processors. In *TeraGrid Conference*, Las Vegas, NV, 2008. http://www.sdsc.edu/us/resources/p3dfft/docs/TG08_DNS.pdf.
- [10] R. A. Gingold and J. J. Monaghan. Smoothed particle hydrodynamics: Theory and application to non-spherical stars. *Mon. Not. R Astron. Soc.*, 181:37–389, 1977.
- [11] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.
- [12] T. Hamada, T. Narumi, R. Yokota, K. Yasuoka, K. Nitadori, and M. Taiji. 42 TFlops hierarchical N-body simulations on GPUs with applications in both astrophysics and turbulence. In *SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, pages 1–12. ACM, 2009.
- [13] T. Hamada and K. Nitadori. 190 TFlops astrophysical N-body simulation on a cluster of GPUs. In *High Performance Computing, Networking, Storage and Analysis (SC), 2010 International Conference for*, pages 1–9, Nov. 2010.
- [14] T. Ishihara, Y. Kaneda, M. Yokokawa, K. Itakura, and A. Uno. Small-scale statistics in high-resolution direct numerical simulation of turbulence: Reynolds number dependence of one-point velocity gradient statistics. *J. Fluid Mech.*, 592:335–366, 2007.
- [15] Christophe G. Lambert, Thomas A. Darden, and John A. Board Jr. A multipole-based algorithm for efficient calculation of forces and potentials in macroscopic periodic assemblies of particles. *J. Comput. Phys.*, 126(2):274–285, 1996.
- [16] I. Lashuk, A. Chandramowlishwaran, H. Langston, T. Nguyen, R. Sampath, A. Shringarpure, R. Vuduc, L. Ying, D. Zorin, and G. Biros. A massively parallel adaptive fast-multipole method on heterogeneous architectures. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC '09*, pages 1–12, Portland, Oregon, November 2009.
- [17] A. Rahimian, I. Lashuk, S. Veerapaneni, A. Chandramowlishwaran, D. Malhotra, L. Moon, R. Sampath, A. Shringarpure, J. Vetter, R. Vuduc, D. Zorin, and G. Biros. Petascale direct numerical simulation of blood flow on 200k cores and heterogeneous architectures. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, SC '10*, pages 1–11. IEEE Computer Society, 2010.
- [18] J. K. Salmon and M. S. Warren. Skeletons from the treecode closet. *J. Comput. Phys.*, 111:136–155, 1994.
- [19] H. Sundar, R. S. Sampath, and G. Biros. Bottom-up construction and 2:1 balance refinement of linear octrees in parallel. *SIAM J. Sci. Comput.*, 30(5):2675–2708, 2008.
- [20] M. S. Warren and J. K. Salmon. Astrophysical N-body simulations using hierarchical tree data structures. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 570–576, Los Alamitos, CA, USA, 1992. IEEE Computer Society Press.
- [21] M. S. Warren and J. K. Salmon. A parallel hashed oct-tree N-body algorithm. In *Proceedings of the 1993 ACM/IEEE Conference on Supercomputing*, pages 12–21, New York, 1993. ACM.
- [22] M. Yokokawa, K. Itakura, A. Uno, T. Ishihara, and Y. Kaneda. 16.4-Tflops direct numerical simulation of turbulence by a Fourier spectral method on the Earth Simulator. In *Supercomputing, ACM/IEEE 2002 Conference*, pages 50–50. IEEE, 2002.
- [23] R. Yokota and L. A. Barba. Hierarchical N-body simulations with auto-tuning for heterogeneous systems. *Computing in Science & Engineering*, 14(3):30–39, May/June 2012.
- [24] R. Yokota and L. A. Barba. Weak scaling of parallel FMM vs. FFT up to 4096 processes. Figure on [figshare](http://dx.doi.org/10.6084/m9.figshare.92425), used under CC-BY license, <http://dx.doi.org/10.6084/m9.figshare.92425>, July 18 2012.
- [25] R. Yokota, R. Sakamaki, S. Kameoka, T. Narumi, K. Yausoka, and S. Obi.

- Fast multipole methods on a cluster of GPUs for the meshless simulation of turbulence. *Comput. Phys. Comm.*, 180(11):2066–2078, 2009.
- [26] R. Yokota, T. K. Sheel, and S. Obi. Calculation of isotropic turbulence using a pure Lagrangian vortex method. *J. Comput. Phys.*, 226:1589–1606, 2007.
 - [27] Rio Yokota and L. A. Barba. Treecode and fast multipole method for N -body simulation with CUDA. In Wen-Mei Hwu, editor, *GPU Computing Gems Emerald Edition*, chapter 9, pages 113–132. Elsevier/ Morgan Kaufman, 2011. Preprint on arXiv:1010.1482.
 - [28] Rio Yokota and L. A. Barba. FMM-based vortex method for simulation of isotropic turbulence on GPUs, compared with a spectral method. To appear in *Comput. & Fluids*, preprint on arXiv:1110.2921, 2012.
 - [29] Rio Yokota, Jaydeep P. Bardhan, Matthew G. Knepley, L. A. Barba, and Tsuyoshi Hamada. Biomolecular electrostatics using a fast multipole BEM on up to 512 GPUs and a billion unknowns. *Comput. Phys. Comm.*, 182(6):1271–1283, 2011.