

High Performance P^3M N-body code: CUBEP3M

Joachim Harnois-Déraps^{1,2} *, Ue-Li Pen¹, Ilian T. Iliev³, Hugh Merz⁴,
J.D. Emberson^{1,5} and Vincent Desjacques⁶

¹Canadian Institute for Theoretical Astrophysics, University of Toronto, M5S 3H8, Ontario, Canada

²Department of Physics, University of Toronto, M5S 1A7, Ontario, Canada

³Astronomy Centre, Department of Physics and Astronomy, Pevensey II Building, University of Sussex, BN1 9QH, Brighton, United Kingdom

⁴SHARCNET, Laurentian University, P3E 2C6, Ontario, Canada

⁵Department of Astronomy and Astrophysics, University of Toronto, M5S 3H4, Ontario, Canada

⁶Université de Genève and Center for Astroparticle Physics, 24 Quai Ernest Ansermet, 1211 Genève 4, Switzerland

30 July 2012

ABSTRACT

This paper presents CUBEP3M, a high performance cosmological N-body code and describes many utilities and extensions that have been added to the standard package, including a run-time halo finder, a non-Gaussian initial conditions generator, a tuneable accuracy, and a system of unique particle identification. CUBEP3M is fast, has a low memory imprint, and has been run on up to 20,000 cores, achieving close to ideal weak scaling even at this scale. It is well suited for a vast number of science applications that require either large samples of non-linear realizations or very large dark matter N-body simulations, including cosmological reionization, baryonic acoustic oscillations, weak lensing or non-Gaussian statistics. We discuss the structure, the accuracy, any known systematic effects, and the scaling performance of the code and its utilities, when applicable.

Key words: N-body simulations — Large scale structure of Universe — Dark matter

1 INTRODUCTION

Many physical and astrophysical systems are subject to non-linear dynamics and rely on N-body simulations to describe the evolution of bodies. One of the main field of application is the modelling of large scale structures, which are driven by the sole force of gravity. Recent observations of the cosmic microwave background (Komatsu et al. 2009, 2011), of galaxy clustering (York et al. 2000; Colless et al. 2003; Schlegel et al. 2009; Drinkwater et al. 2010) of weak gravitational lensing (Heymans & CFHTLenS Collaboration 2012; Sheldon et al. 2009) and of supernovae redshift-distance relations all point towards a standard model of cosmology, in which dark energy and collision-less dark matter occupy more than 95 per cent of the total energy density of the universe. In such a paradigm, pure N-body code are perfectly suited to describe the dynamics, as long as baryonic physics is not very important, or at least we understand how the baryonic fluid feeds back on the dark matter structure. The next generation of measurements aim at constraining the cosmological parameters at the per cent level, and the theoretical understanding of the non-linear dynamics that govern structure formation heavily relies on numerical simulations.

For instance, a measurement of the baryonic acoustic oscillation (BAO) dilation scale can provide tight constraints on the dark energy equation of state (Eisenstein et al. 2005; Tegmark et al.

2006; Percival et al. 2007; Schlegel et al. 2009). The most optimal estimates of the uncertainty requires the knowledge of the matter power spectrum covariance matrix, which is only accurate when measured from a large sample of N-body simulations (Rimes & Hamilton 2005; Takahashi et al. 2009, 2011). For the same reasons, the most accurate estimates of weak gravitational lensing signal is obtained by propagating photons in past light cones that are extracted from simulated density fields (Vale & White 2003; Sato et al. 2009; Hilbert et al. 2009). Another area where large-scale N-body simulations have in recent years been instrumental are in simulations of early cosmic structures and reionization (e.g. Iliev et al. 2006; Zahn et al. 2007; Trac & Cen 2007; Iliev et al. 2011). The reionization process is primarily driven by low-mass galaxies, which for both observational and theoretical reasons, need to be resolved in fairly large volumes, which demands simulations with a very large dynamic range.

The basic problem that is addressed with N-body codes is the time evolution of an ensemble of N particles that is subject to gravitational attraction. The brute force calculation requires $O(N^2)$ operations, a cost that exceeds the memory and speed of current machines for large problems. Solving the problem on a mesh (Hockney & Eastwood 1981) reduces to $O(N \log N)$ the number of operations, as it is possible to solve for the particle-mesh (PM) interaction with fast Fourier transforms techniques with high performance libraries such as FFTW (Frigo & Johnson 2005).

With the advent of large computing facilities, parallel compu-

* E-mail: jharno@cita.utoronto.ca

tations have now become common practice, and N-body codes have evolved both in performance and complexity. Many have opted for ‘tree’ algorithms, including GADGET (Springel et al. 2001; Springel 2005), PMT (Xu 1995), GOTPM (Dubinski et al. 2004), and Hydra (Couchman et al. 1995), in which the local resolution increases with the density of the matter field. These often have the advantage to balance the work load across the computing units, which enable fast calculations even in high density regions. The drawback is a significant loss in speed, which can be only partly recovered by turning off the tree algorithm. The same reasoning applies to mesh-refined codes (Couchman 1991), which in the end are not designed to perform fast PM calculations on large scales.

Although such codes are needed to study systems that are spatially heterogeneous like individual clusters or haloes, AGNs or other compact objects, many applications are interested in studying large cosmological volumes in which the matter distribution is rather homogeneous. In such environments, the load balancing algorithm can be removed, and the effort can thereby be put towards speed, memory compactness and scalability. PMFAST (Merz et al. (2005), MPT hereafter) was one of the first code designed specifically to optimize the PM algorithm, both in terms of speed and memory usage, and uses a two-level mesh algorithm based on the gravity solver of Trac & Pen (2003). The long range gravitational force is computed on a grid 4^3 times coarser, such as to minimize the MPI communication time and to fit in system’s memory. The short range is computed locally on a finer mesh, and only the local sub-volume needs to be stored at a given time, allowing for OPENMP parallel computation. This setup enables the code to evolve large cosmological systems both rapidly and accurately, even on relatively modest clusters. One of the main advantages of PMFAST over other PM codes is that the number of large arrays is minimized, and the global MPI communications are cut down to the minimum: for passing particles at the beginning of each time step, and for computing the long range FFTs.

Since its first published version, PMFAST has evolved in many aspects. The first major improvement was to transform the volume decomposition in multi-node configurations from slabs to cubes. The problem with slabs is that they do not scale well to large runs: as the number of cells per dimension increases, the thickness of each slab shrinks rapidly, until it reaches the hard limit of a single cell layer. With this enhancement, the code name was changed to CUBEP3M. Soon after, it incorporated particle-particle (pp) interactions at the sub-grid level, and was finally renamed CUBEP3M. The public package now includes a significant number of new features: the pp force can be extended to an arbitrary range, the size of the redshift jumps can be constrained for improved accuracy during the first time steps, a runtime halo finder has been implemented, the expansion has also been generalized to include a redshift dependent equation of state of dark energy, there is a system of unique particle identification that can be switched on or off, and the initial condition generator has been generalized as to include non-Gaussian features. PMFAST was equipped with a multi-time stepping option that has not been tested on CUBEP3M yet, but which is, in principle at least, still available.

The standard package also contains support for gas cosmological evolution through a portable TVD-MHD module (Pen et al. 2003) that scales up to thousands of cores as well (see Pang et al. (2010) and footnote 4 in section 4), and a coupling interface with the radiative transfer code C2-Ray (Mellema et al. 2006). CUBEP3M is therefore one of the most competitive and versatile public N-body code, and has been involved in a number of scientific applications over the last few years, spanning the field of

weak lensing (Vafaei et al. 2010; Lu & Pen 2008; Doré et al. 2009; Lu et al. 2010; Yu et al. 2010; Harnois-Déraps et al. 2012), BAO (Zhang et al. 2010; Ngan et al. 2012; Harnois-Déraps & Pen 2012; Harnois-Déraps et al. 2012), formation of early cosmic structures (Iliev et al. 2008, 2010), observations of dark stars (Zackrisson et al. 2010; Ilie et al. 2012) and reionization (Iliev et al. 2011; Fernandez et al. 2011; Friedrich et al. 2011; Mao et al. 2012; Datta et al. 2011; Friedrich et al. 2012). Continuous efforts are being made to develop, extend and improve the code and each of its utilities, and we expect that this will pave the way to an increasing number of science projects. Notably, the fact that the force calculation is multi-layered makes the code extendible, and opens the possibility to run CUBEP3M on hybrid CPU-GPU clusters. It is thus important for the community to have access to a paper that describes the methodology, the accuracy and the performance of this public code.

Since CUBEP3M is not new, the existing accuracy and systematic tests were performed by different groups, on different machines, and with different geometric and parallelization configurations. It is not an ideal situation in which to quantify the performance, and each test must be viewed as a separate measurement that quantifies a specific aspect of the code. We have tried to keep to a minimum the number of such different runs, and although the detailed numbers vary with the problem size and the machines, the general trends are rather universal.

Tests on constraints of the redshift step size and on improvements of the force calculations were performed on the Sunnyvale beowulf cluster at the Canadian Institute for Theoretical Astrophysics (CITA). Each node contains 2 Quad Core Intel(R) Xeon(R) E5310 1.60GHz processors, 4GB of RAM, a 40 GB disk and 2 gigE network interfaces. Each of these tests were performed on the same cluster, but with different box sizes, starting redshifts and particle numbers. Hereafter we refer to these simulation sets as the CITA configurations, and specify which parameters were used in each case. Some complimentary runs were also performed on the SciNet GPC cluster (Loken et al. 2010).

For tests of the code accuracy, of the non-Gaussian initial conditions generator and of the run time halo finder algorithm, we used a third simulation configuration series that was run at the Texas Advanced Computing Centre (TACC) on Ranger, a Sun-Blade x6420 system with AMD x86_64 Opteron Quad Core 2.3 GHz ‘Barcelona’ processors and Infiniband networking. These RANGER4000 simulations evolved 4000^3 particles from $z = 100$ to $z = 0$ with a box side of $3.2h^{-1}\text{Gpc}$ on 4000 cores.

The paper is structured as follow: section 2 reviews the structure and flow of the main code; section 3 describes how Poisson equation is solved on the two-mesh system; we then present in section 4 the scaling of the code to very large problems, including much larger runs that were produced at various high performance computing centers; section 5 discuss the accuracy and systematic effects of the code. We then describe in section 6 the run-time halo finder, in section 7 various extensions to the default configuration, and conclude afterwards.

2 REVIEW OF THE CODE STRUCTURE

An optimal large scale N-body code must address many challenges: minimize the memory footprint to allow larger dynamical range, minimize the passing of information across computing nodes, reduce and accelerate the memory accesses to the large scale arrays, make efficient use of high performance libraries to speed up standard calculations like Fourier transforms, just to name a few. In

the realm of parallel programming, high efficiency can be assessed when a high load is balanced across all processors most of the time. In this section, we present the general strategies adopted to address these challenges¹. We start with a walkthrough the code flow, and briefly discuss some specific sections that depart from standard N-body codes, while referring the reader to future sections for detailed discussions on selected topics.

As mentioned in the Introduction section, CUBEP3M is a FORTRAN90 N-body code that solves Poisson's equation on a two-level mesh, with sub-cell accuracy thanks to particle-particle interactions. The code has extensions that departs from this basic scheme, and we shall come back to these later, but for the moment, we adopt the standard configuration. The long range component of the gravity force is solved on the coarse grid, and is global in the sense that the calculations require knowledge about the full simulated volume. The short range force and the particle-particle interactions are computed in parallel on a second level of cubical decomposition of the local volumes, the *tiles*. To make this possible, the fine grid arrays are constructed such as to support parallel memory access. In practice, this is done by adding an additional dimension to the relevant arrays, such that each CPU accesses a unique memory location. The force matching between the two meshes is performed by introducing a cutoff length, r_c , in the definition of the two force kernels. The value of $r_c = 16$ fine cells was found to balance the communication overhead between processes and the accuracy of the match between the two meshes.

The computation of the short range force requires each tile to store the fine grid density of a region that includes a buffer surface around the physical volume it is assigned. The thickness of this surface must be larger than r_c , and we find that a 24 cells deep buffer is a good compromise between memory usage and accuracy. This is basically to fully compensate for the coarse mesh calculations, whose CIC interpolation scheme reaches two coarse cells deep beyond the cutoff.

When it comes to finding haloes at run time, this buffer can create a problem, because a large object located close to the boundary can have a radius larger than the buffer zone, in which case it would be truncated and be assigned a wrong mass, center of mass, etc. It could then be desirable to increase the buffer zone around each tile, at the cost of a loss of memory dedicated to the actual physical volume, and the code is designed to allow for such a flexibility.

Because the coarse grid arrays require 4^3 times less memory per node, it does not contribute much to the total memory requirement, and the bulk of the foot-print is concentrated in a handful of fine grid arrays. Some of these are only required for intermediate steps of the calculations, hence it is possible to hide therein many coarse grid arrays². We present here the largest arrays used by the code:

- (i) `xv` stores the position and velocity of each particle
- (ii) `ll` stores the linked-list that accelerate the access to particles in each coarse grid cell
- (iii) `rho_f` and `cmplx_rho_f` store the local fine grid density in real and Fourier space respectively
- (iv) `force_f` stores the force of gravity (short range only) along the three Cartesian directions

(v) `kern_f` stores the fine grid force kernel in the three directions

(vi) `PID` stores the unique particle identification tags.

The particle ID is a feature that can be switched off by removing a compilation flag, and allows to optimize the code for higher resolution configurations. In terms of memory, CUBEP3M can be configured such as the main footprint is dominated exclusively by the particle phase space array, in which case it uses about 60 bytes per particles. For a fixed problem size, this can be achieved by increasing the number of tiles, thereby reducing the size of the local fine mesh arrays. There is a limit at which we can do such a volume breakdown, though, since the physical volume on each tile must be much larger than the connection length between the two meshes (see section 3). For comparison, (Lean) GADGET-2 code uses around 90 per particles.

The code flow is presented in Fig. 1 and 2. Before entering the main loop, the code starts with an initialization stage, in which many declared variables are assigned default values, the redshift checkpoints are read, the FFTW plans are created, and the MPI communicators are defined. The phase-space array is obtained from the output of the initial conditions generator, and the force kernels on both grids are constructed by reading precomputed kernels which are then adjusted to the specific simulation size. For clarity, all these operations are collected under the subroutine call `initialize` in Fig. 1, although they are actually distinct calls in the code.

As described in MPT, access to particles is accelerated with the use of linked lists, deletion of 'ghost' particles in buffer zones is done at the same time as particles are passed to adjacent nodes, and the global FFTs are performed with a slab decomposition of the volumes via a special file transfer interface, designed specifically to preserve a high processor load.

Each iteration of the main loop starts with the `timestep` subroutine, which proceeds to a determination of the redshift jump by comparing the step size constraints from each force components and from the scale factor. The cosmic expansion is found by Taylor expanding Friedmann's equation up to the third order in the scale factor, and can accommodate constant or running equation of state of dark energy. The force of gravity is then solved in the `particle_mesh` subroutine, which first updates the positions and velocities of the dark matter particles, exchange with neighbouring nodes those that have exited to volume, creates a new linked list, then solve Poisson's equation. This subroutine is conceptually identical to that of PMFAST, with the exception that CUBEP3M decomposes the volume into cubes (as opposed to slabs). The loop over tiles and the particle exchange are thus performed in three dimensions. When the short range and pp forces have been calculated on all tiles, the code exits the parallel OPENMP loop and proceeds to the long range. This section of the code is also parallelized on many occasions, but, unfortunately, the current MPI-FFTW do not allow multi-threading. There is thus an inevitable loss of efficiency during each global Fourier transforms, during which only the single core MPI process is active³. As in PMFAST, the particle position and velocity updates are performed in a leap frog scheme (Hockney & Eastwood 1981).

If the current redshift corresponds to one of the checkpoints, the code advances all particles to their final location and writes

¹ Many originate directly from MPT and were preserved in CUBEP3M; those will be briefly mentioned, and we shall refer the reader to the original PMFAST paper for greater details.

² This memory recycling is done with 'equivalence' statements in F90

³ Other libraries such as P3DFFT (<http://code.google.com/p/p3dfft/>) currently permit an extra level of parallelization, and it is our plan to migrate to one of these in the near future.

```

program cubep3m
  call initialize
  do
    call timestep
    call particle_mesh
    if(checkpoint_step) then
      call checkpoint
    elseif(last_step)
      exit
    endif
  enddo
  call finalize
end program cubep3m

```

Figure 1. Overall structure of the code simplified for readability.

```

subroutine particle_mesh
  call update_position + apply random offset
  call link_list
  call particle_pass
  !$omp parallel do
  do tile = 1, tiles_node
    call rho_f_ngp
    call cmplx_rho_f
    call kernel_multiply_f
    call force_f
    call update_velocity_f
    if(pp = .true.) then
      call link_list_pp
      call force_pp
      call update_velocity_pp
      if(extended_pp = .true.) then
        call link_list_pp_extended
        call force_pp_extended
        call update_velocity_pp_extended
      endif
    endif
  enddo
  !$omp end parallel do
  call rho_c_ngp
  call cmplx_rho_c
  call kernel_multiply_c
  call force_c
  call update_velocity_c
  delete_buffers
end subroutine particle_mesh

```

Figure 2. Overall structure of the two-level mesh algorithm. We have included the section that concerns the standard pp and the extended pp force calculation, to illustrate that they follow similar linked-list logic.

them to file. These restart files start with a small header that contains the local number of particles, the current redshift and time step, and the constraints on the time step jump from the previous iteration; the header is followed by the phase space of each particles. The particle identification tags are written with a similar fashion in distinct files to simplify the post-processing coding. This general I/O strategy allows for highly efficient MPI parallel read and write, by default in binary format for compactness, and has been shown to scale well to large data sets.

Similarly, the code can compute two-dimensional projections of the density field, halo catalogues (see section 6 for details), and can compute the power spectrum on the coarse grid at run time. The code exits the loop when it has reached the final redshift, it then wraps up the FFTW plans and clears the MPI communicators.

We have collected those operations under the subroutine `finalize` for concision.

Other constraints that need to be considered is that any decomposition geometry somehow limits permissible grid sizes, and that volume evenly decomposed into cubic sub-sections – such as CUBEP3M – requires the number of MPI processes to be a perfect cube. Also, since the decomposition is volumetric, as opposed to density dependent – it suffers from load imbalance whenever the particles are not evenly distributed. However, large scale cosmology problems are very weakly affected by this effect, which can generally be overcome by allocating a mild amount of extra memory per node. As mentioned in section 4, large runs can even be optimized by allocating only a small amount for the first part of it, follow by a restart with a configuration that allows more memory per node.

3 POISSON SOLVER

This section reviews how Poisson’s equation is solved on a double-mesh configuration. Many parts of the algorithm are identical to PMFAST, hence we refer the reader to section 2 of MPT for more details. In CUBEP3M, the mass default assignment scheme are a ‘cloud-in-cell’ (CIC) interpolation for the coarse grid, and a ‘nearest-grid-point’ (NGP) interpolation for the fine grid (Hockney & Eastwood 1981). This choice is motivated by the fact that the most straightforward way to implement a P³M algorithm on a mesh is to have exactly zero mesh force inside a grid, which is only true for the NGP interpolation. Although CIC generally has a smoother and more accurate force, the pp implementation enhances the code resolution by almost an order of magnitude.

The code units inherit from (Trac & Pen 2004) and are summarized here for completeness. The comoving length of a fine grid cell is set to one, such that the unit length in simulation unit is

$$1\mathcal{L} = a \frac{L}{N} \quad (1)$$

where a is the scale factor, N is the total number of cells along one dimension, and L is the comoving volume in $h^{-1}\text{Mpc}$. The mean comoving mass density is also set to unity in simulation units, which, in physical units, corresponds to

$$1\mathcal{D} = \rho_m(0)a^{-3} = \Omega_m \rho_c a^{-3} = \frac{3\Omega_m H_o^2}{8\pi G a^3} \quad (2)$$

Ω_m is the matter density, H_o is the Hubble’s constant, ρ_c is the critical density today, and G is Newton’s constant. The mass unit is found with $\mathcal{M} = \mathcal{D}\mathcal{L}^3$. Specifying the value of G on the grid fixes the time unit, and with $G_{\text{grid}} = 1/(6\pi a)$, we get:

$$1\mathcal{T} = \frac{2a^2}{3} \frac{1}{\sqrt{\Omega_m H_o^2}} \quad (3)$$

These choices completely determine the conversion between physical and simulation units. For instance, the velocity units are given by $1\mathcal{V} = \mathcal{L}/\mathcal{T}$.

The force of gravity on a mesh can be computed either with a gravitational potential kernel $\omega_\phi(\mathbf{x})$ or a force kernel $\omega_F(\mathbf{x})$. Gravity fields are curl-free, which allows us to relate the potential $\phi(\mathbf{x})$ to the source term via Poisson’s equation:

$$\nabla^2 \phi(\mathbf{x}) = 4\pi G \rho(\mathbf{x}) \quad (4)$$

We solve this equation in Fourier space, where we write:

$$\tilde{\phi}(\mathbf{k}) = \frac{-4\pi G \tilde{\rho}(\mathbf{k})}{k^2} \equiv \tilde{\omega}_\phi(\mathbf{k}) \tilde{\rho}(\mathbf{k}) \quad (5)$$

The potential in real space is then obtained with an inverse Fourier transform, and the kernel becomes $\omega_\phi(\mathbf{x}) = -G/r$. Using the convolution theorem, we can write

$$\phi(\mathbf{x}) = \int \rho(\mathbf{x}') \omega_\phi(\mathbf{x}' - \mathbf{x}) d\mathbf{x}' \quad (6)$$

and

$$\mathbf{F}(\mathbf{x}) = -m\nabla\phi(\mathbf{x}) \quad (7)$$

Although this approach is fast, it involves a finite differentiation at the final step, which enhances the numerical noise. We therefore opt for a force kernel, which is more accurate, even though it requires four extra Fourier transforms. In this case, we must solve the convolution in three dimensions and define the force kernel ω_F such as:

$$\mathbf{F}(\mathbf{x}) = \int \rho(\mathbf{x}') \omega_F(\mathbf{x}' - \mathbf{x}) d\mathbf{x}' \quad (8)$$

Because the gradient acting on 6 affects only unprime variables, we can express the force kernel as a gradient of the potential kernel. Namely:

$$\omega_F(\mathbf{x}) \equiv -\nabla\omega_\phi(\mathbf{x}) = -\frac{mG\hat{\mathbf{r}}}{r^2} \quad (9)$$

Following the spherically symmetric matching technique of MPT (section 2.1), we split the force kernel into two components, for the short and long range respectively, and match the overlapping region with a polynomial. Namely, we have:

$$\omega_s(r) = \begin{cases} \omega_F(r) - \beta(r) & \text{if } r \leq r_c \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

and

$$\omega_l(r) = \begin{cases} \beta(r) & \text{if } r \leq r_c \\ \omega_F(r) & \text{otherwise} \end{cases} \quad (11)$$

The vector $\beta(r)$ is related to the fourth order polynomial that is used in the potential case described in MPT by $\beta = -\nabla\alpha(r)$. The coefficients are found by matching the boundary conditions at r_c up to the second derivative, and we get

$$\beta(r) = \left[-\frac{7r}{4r_c^3} + \frac{3r^3}{4r^5} \right] \hat{\mathbf{r}} \quad (12)$$

Since these calculations are performed on two grids of different resolution, a sampling window function must be convoluted both with the density and the kernel (see [Eq. 7-8] of MPT). When matching the two force kernels, the long range force is always on the low side close to the cutoff region, whereas the short range force is uniformly scattered across the theoretical $1/r^2$ value – intrinsic features of the CIC and NGP interpolation schemes respectively. By performing force measurements on two particles randomly placed in the volume, we identified a small region surrounding the cutoff length in which we empirically adjust both kernels such as to improve the match. Namely, for $14 \leq r \leq 16$, $\omega_s(r) \rightarrow 0.985\omega_s(r)$, and for $12 \leq r \leq 16$, $\omega_l(r) \rightarrow 1.2\omega_l(r)$.

As mentioned in section 2 and summarized in Fig. 2, the force kernels are first in the code initialization stage. Eq. 8 is then solved with fast Fourier transforms along each direction, and is applied onto particles in the `update_velocity` subroutine.

The pp force is calculated during the fine mesh velocity update, which avoids loading the particle list twice and allows the operation to be threaded without significant additional work. During this process, the particles within a given fine mesh tile are first

Table 1. Scaling in CPU resources as a function of the softening length.

r_{soft}	time (h)
0.5	5.75
0.3	8.17
0.2	10.09
0.1	18.67
0.05	31.07

read in via the linked list, then their velocity is updated with the fine mesh force component, according to their location within the tile. In order to organize the particle-particle interactions, we proceed in constructing a set of threaded fine-cell linked list chains for each coarse cell. We then calculate the pairwise force between all particles that lie within the same fine mesh cell, excluding pairs whose separation is smaller than a softening length r_{soft} : particles separated by less than this distance have their particle-particle force set to zero. As this proceeds, we accumulate the pp force applied on each particle and then determine the maximum force element of the pp contribution, which is also taken into account when constraining the length of the global time step.

Force softening is generally required by any code to prevent large scattering as $r \rightarrow 0$ that can otherwise slow the calculation down, and to reduce the two-body relaxation, which can affect the numerical convergence. Many force softening schemes can be found in the literature, including Plummer force, uniform or linear density profiles or the spline-softened model (see Dyer & Ip (1993) for a review). In the current case, a sharp force cutoff corresponds to a particle interacting with a hollow shell. In comparison with other techniques, this force softening is the easiest to code and the fastest to execute. Generally, it is desirable to match the smoothness of the force to the order of the time integration. A Plummer force is infinitely differentiable, which is a sufficient but not necessary condition for our 2^{nd} order time integration. Also, one of the drawbacks of Plummer's softening is that the resolution degrades smoothly: the effects of the smoothing are present at all radii. In comparison, the uniform density and hollow shell alternatives both have the advantage that the deviations from $1/r^2$ are minimized. Although all the results presented in this paper were obtained with the sharp cutoff softening, other schemes can easily be adopted as these are typically single line changes to the code. Dyer & Ip (1993) argues in favour on the uniform density profile scheme – which is more physical – and future developments of CUBEP3M will incorporate this option.

The choice of softening length is motivated by a trade off between accuracy and run time. Larger values reduce the structure formation but run quicker. We show in Fig. 3 the impact of changing this parameters on the power spectrum. For this test, we produced SciNet256 simulations, starting at $z = 100$ and evolving to $z = 0.5$, with a box size of $100h^{-1}\text{Mpc}$. We also record in Table 1 the real run time for each trials, from where we see the strong effect of computation time the softening length has. In this test, which is purposefully probing rather deep in the non-linear regime, reducing the softening length to half its value ($1/20^{th}$ of a grid cell) double the run time. The gain in power spectrum is less than two per cent. Similarly, increasing to 0.2 the softening length reduces almost by a half the run time, but suffers from a five per cent loss in power at the resolution turn-around. One tenth of grid cell seems to be the optimal choice in this trade off, and is therefore the default value, however it should really be considered a free parameter.

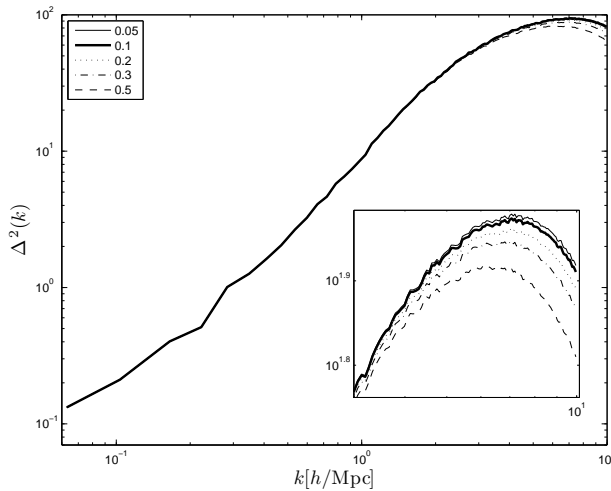


Figure 3. Dark matter power spectrum, measured at $z = 0.5$ in a series of SciNet256 simulations in which the softening length is varied. The simulations started off at $z = 100$, and have a box size of $100h^{-1}\text{Mpc}$.

In addition, PMFAST could run with a different set of force kernels, described in MPT as ‘least square matching’, which basically adjust the kernels on cell-by-cell basis based on minimization of the deviation with respect to Newtonian predictions. This was originally computed such as to optimize the force calculation for the case where both grids are obtained from CIC interpolation. Moving to a mix CIC/NGP scheme requires solving the system of equations with the new configuration, a straightforward operation. With the inclusion of the random shifting, it is not clear how much improvement one would recover from this other kernel matching. It is certainly something we will investigate and document in the near future.

Finally, a choice must be done concerning the longest range of the coarse mesh force. Gravity can be either a) an accurate $1/r^2$ force, as far as the volume allows, or b) modified to correctly match the periodicity of the boundary conditions. By default, the code is configured along to the second choice, which accurately models the growth of structures at very large scales. However, detailed studies of gravitational collapse of a single large object would benefit from the first setting, even though the code is not meant to evolve such systems that generally require load balance control.

4 SCALING PERFORMANCES

The parallel algorithm of CUBEP3M is designed for ‘weak’ scaling, i.e. if the number of cores and the problem size increase in proportion to each other, then for ideal scaling the wall-clock time should remain the same. This is to be in contrasted with ‘strong’ scaling codes, whereby the same problem solved on more cores should take proportionately less wall-clock time. This weak scaling requirement is dictated by the problems we are typically investigating (very large and computationally-intensive) and our goals, which are to address such large problems in the most efficient way, rather than for the least wall-clock time. Furthermore, we recall that there is no explicit load balancing feature, thus the code is maximally efficient when the sub-domains contain roughly an equal number of particles. This is true for most cosmological-size volumes that do not resolve too deep in the non-linear regime, but not for e.g. simulations of a single highly-resolved galaxy.

Because of the volumetric decomposition, the total number of MPI processes needs to be a perfect cube. Also, for maximal resource usage, the number of tiles per node should be a multiple of the number of available CPUs per MPI process, such that no core sits idle in the threaded block. Given the available freedom in the parallel configuration, as long as the load is balanced, it is generally good practice to maximize the number of OPENMP threads and minimize the number of MPI processes: the information exchange between cores that are part of the same motherboard is generally much faster. In addition, having fewer MPI processes reduces the total amount of buffer zones, freeing memory that can be used to increase the mesh resolution. However, it has been observed that for the case of non-uniform memory access (NUMA) systems, the code is optimized when only the cores that share the same socket should be OPENMP threaded. As one probes deeper into the non-linear regime however, the formation of dense objects can cause memory problems in such configurations, and increasing the number of MPI processes helps to ensure memory locality, especially in NUMA environments.

The intermediary version of the code – CUBEPM – was first ported to the IBM Blue Gene/L platform, and achieved weak-scaling up to 4096 processes (over a billion particles), with the N-body calculation only incurring a 10 per cent overhead at runtime (compared to 8 processes) for a balanced workload⁴. In order to accommodate the limited amount of memory available per processing core on the Blue Gene/L platform machines, it was necessary to perform the long range MPI FFT with a volumetric decomposition (Eleftheriou et al. 2005). Slab decomposition would have required a volume too large to fit in system memory given the constraints in the simulation geometry.

The scaling of CUBEP3M was first tested with a dedicated series of simulations – the CURIE simulation suite – by increasing the size and number of cores on the ‘fat’ (i.e. large-memory) nodes of the Curie supercomputer at the Très Grand Centre de Calcul (TGCC) in France. For appropriate direct comparison, all these simulations were performed using the same particle mass ($M_{\text{particle}} = 1.07 \times 10^{11} M_{\odot}$) and force resolution (softening length $50 h^{-1}\text{kpc}$). The box sizes used range from $256 h^{-1}\text{Mpc}$ to $2048 h^{-1}\text{Mpc}$, and the number of particles from 256^3 to 2048^3 . Simulations were run on 32 up to 2048 computing cores, also starting from redshift $z = 100$, and evolving until $z = 0$. Our results are shown in Fig. 4 and in Table 2, and present excellent scaling, within ~ 3 per cent of ideal, at least for up to 2048 cores.

We have also ran CUBEP3M on a much larger number of cores, from 8000 to up to 21,976, with 5488^3 – 6000^3 (165 to 216 billion) particles on Ranger and on JUROPA at the Jülich Supercomputing Centre in Germany, which is an Intel Xeon X5570 (Nehalem-EP) quad-core 2.93 GHz system, also interconnected with Infiniband. Since it is not practical to perform dedicated scaling tests on such a large number of computing cores, we instead list in Table 3 the data directly extracted from production runs. We have found the code to scale within 1.5 per cent of ideal up to 4096 cores. For larger sizes ($\geq 10,976$ cores), the scaling is less ideal, due to increased communication costs, I/O overheads (a single timeslice of 5488^3 particles is 3.6 TB) and load balancing issues, but still within ~ 20

⁴ http://web.archive.org/web/20060925132146/http://www-03.ibm.com/servers/deepcomputing/pdf/Blue_Gene_Applications_Paper_CubePM_032306.pdf

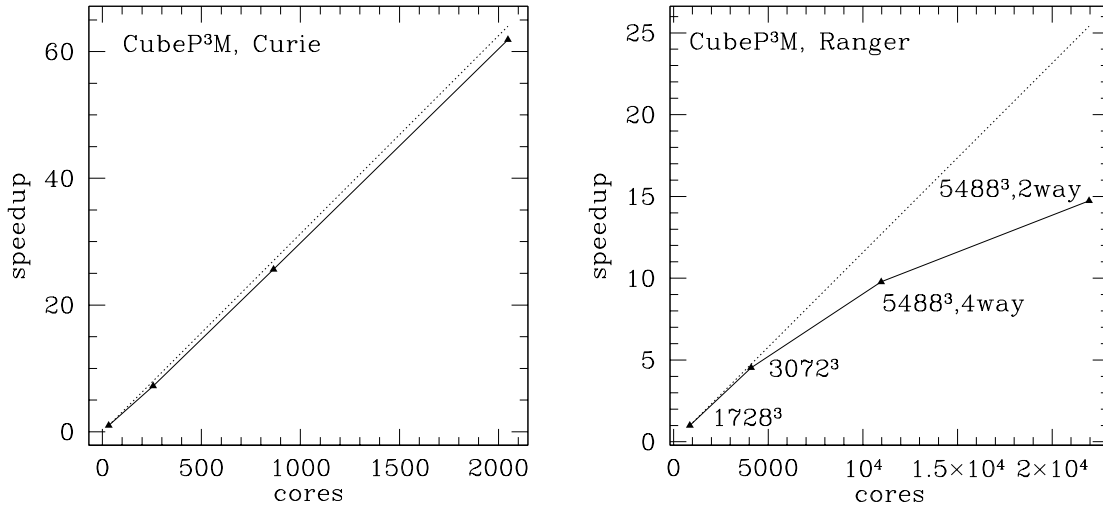


Figure 4. Scaling of CUBEP3M on Curie fat nodes (left) and on Ranger TACC facility for very large number of cores (right). Plotted is the code speedup ($N_{\text{particles}}^3/t_{\text{wallclock}}$) against core count, normalized by the smallest run in each case. Dashed line indicates the ideal weak scaling. The data are listed in Table 2.

Table 2. Scaling of CUBEP3M on Curie. Speedup is scaled to the smallest run.

number of cores	speedup	ideal speedup	absolute timing (min)	$N_{\text{particles}}$	box size (h^{-1} Mpc)
32	1.00	-	3.2	256^3	256
256	7.21	8	3.55	512^3	512
864	25.63	27	4.8	864^3	864
2048	61.87	64	26.48	2048^3	2048

Table 3. Scaling of CUBEP3M on Ranger. Speedup is scaled to the smallest run.

number of cores	speedup	ideal speedup	absolute timing (min)	$N_{\text{particles}}$	box size (h^{-1} Mpc)
864	1.00	-	258	1728^3	6.3
4096	4.53	4.74	320	3072^3	11.4
10976	9.78	12.7	845	5488^3	20
21952	14.73	25.4	561	5488^3	20

per cent of ideal. These first three Ranger runs were performed with 4 MPI processes and 4 threads per Ranger node ('4way')⁵.

Furthermore, due to the increasing clustering of structures at those small scales, some of the cuboid sub-domains came to contain a number of particles well above the average, thereby requiring more memory per MPI process in order to run until the end. As a consequence, throughout most of their late evolution, the largest two of these simulations were run with 4096 and 21,952 cores and with only 2 MPI processes and 8 threads per Ranger node ('2way'), which on Ranger allows using up to 16 GB of RAM per MPI process⁶. Because each processor accesses memory that is not fully

local, this configuration does affect the performance somewhat, as does the imperfect load balancing that arises in such situations. This can be seen in the rightmost point of Fig. 4 (right panel), where the scaling is 42 per cent below the ideal. We note that we still get ~ 1.5 speedup from doubling the core count, even given these issues. Overall the code scaling performance is thus satisfactory even at extremely large number of cores. We expect the code to handle even larger problems efficiently, and is thus well suited to run on next generation Petascale systems.

Finally, we note that several special fixes had to be developed by the TACC and JUROPA technical staff in order for our largest runs to work properly. In particular, we encountered unexpected problems from software libraries such as MPICH and FFTW when applied to calculations of such unprecedented size.

5 ACCURACY AND SYSTEMATICS

This section describes the systematics effects that are inherent to our P^3M algorithm. We start with a demonstration of the accuracy with a power spectrum measurement of a RANGER4000 simula-

⁵ For these very large runs, we used a NUMA script *tacc_affinity*, specially-provided by the technical staff, that bind the memory usage to local sockets, thus ensuring memory affinity. This becomes important because the memory sockets per node (32 GB RAM/node on Ranger) are actually not equal-access. Generally, the local memory of each processor has much shorter access time.

⁶ In order to insure local memory affinity, a second special NUMA control script, *tacc_affinity_2way*, was developed for us by the TACC technical staff and allowed to run more efficiently in this mode.

tion. The halo mass function also assess the code capacity to model gravitational collapse, but depends on the particular halo finder used. We thus post-pone the discussion on this aspect until section 6, and focus for the moment on the particles only. In addition to the power spectrum, we further quantify the accuracy of the force calculation with comparisons to Newton's law of gravitation.

5.1 Density and Power Spectrum

One of the most reliable ways to assess the simulation's accuracy at evolving particles is to measure the density power spectrum at late time, and compare to non-linear prediction. For an over-density field $\delta(x)$, the power spectrum is extracted from the two point function in Fourier space as:

$$\langle |\delta(k)\delta(k')| \rangle = (2\pi)^3 P(k)\delta_D(k' - k) \quad (13)$$

where the angle bracket corresponds to an ensemble (or volume) average. Fig. 5 presents a 2-dimensional density projection of a RANGER4000 simulation, which evolved 4000^3 particles until redshift zero. We then plot in Fig. 6 its dimensionless power spectrum and observe that the agreement with the non-linear prediction of Smith et al. (2003) is within five per cent up to $k = 1.0h\text{Mpc}^{-1}$. The code exhibits a ~ 5 per cent over estimate compared to the theory for $0.2 < k < 0.6h\text{Mpc}^{-1}$, an effect generally caused by inaccuracies at the early time steps. It can be removed by starting the code at a later redshift, but then the code has less time to relax, and other quantities like the halo mass functions or the four point functions become less accurate. The drop of power at higher k is partly caused by the finite mesh resolution, partly from expected deviations about the predictions. The fluctuations at low- k come from the white noise imposed in the initial conditions, and it was show in Ngan et al. (2012) and Harnois-Déraps & Pen (2012) that samples of a few hundreds of realizations average to the correct value.

5.2 Mesh force at grid distances

The force test presented in MPT was carried by placing two particles at random locations on the grid, calculating the force between them, then iterating over different locations. This pairwise force test is useful to quantify the accuracy on a cell-by-cell basis, but lacks the statistics that occur in an actual time step calculation. The actual force of gravity in the P³M algorithm, as felt by a single particle during a single step, is presented in the top left panel of Fig. 7. This force versus distance plot was obtained from a CITA128 realization, and the calculation proceeds in two steps: 1- we compute the force on each particle in a given time step. 2- we remove a selected particle, thus creating a 'hole', compute the force again on all particles, and record on file the force difference (before and after the removal) as a function of the distance to the hole.

Particles in the same fine cell as the hole follow the exact $1/r^2$ curve. The scatter at distances of the order of the fine grid is caused by the NGP interpolation scheme: particles in adjacent fine cells can be actually very close, as seen in the upper left region of this plot, but still feel the same mesh force at grid cell distances. This creates a discrepancy up to an order of magnitude, in loss or gain, depending on the location of the pair with respect to the centre of the cell. As the separation approaches a tenth of the full box size or so, the force on the coarse mesh scatters significantly about Newton's law due to periodic boundary conditions. As mentioned at the end of section 3, the longest range of force kernel can either model accurately Newtonian gravity, or the structure growth of the largest

modes, but not both. For the sake of demonstrating the accuracy of the force calculation, we chose the first option in this section, but this is not how the code would normally be used.

The top right panel of Fig. 7 shows the fractional error on the force along the radial direction (top) and the fractional tangential contribution (bottom), also calculated from a single time step. The fractional error of the radial component is up to 60 per cent at sub-grid distances, but remains under ten per cent for distances larger than a few cells. The transverse fractional error peaks exactly at the grid distance, and is about two time smaller than the radial counterpart.

Although these scatter plots show the contributions from individual particles and cells, it is not clear whether the mean radial force felt is higher or lower the predictions, hence we rebin the results in 50 logarithmically spaced bins and compute the mean and standard deviation. We show this plot in the middle left panel of Fig. 7, where we observe that on average, there is a loss of force of the order of unity in the range $0.4 < r < 1$ (in units of fine grid), but the numerical calculations are otherwise in excellent agreement with Newton's law. The transverse force is somehow ill-defined, since it is by definition a positive number. Since we know that it averages out to zero over many time steps, we plot only the scatter about the mean, represented in the figure by the solid line.

At grid scales, the fractional error is larger than PMFAST, largely due to the fact that the fine mesh force is performed with an NGP interpolation scheme – as opposed to CIC. This prescription is responsible for the larger scatter about the theoretical value, but, as mentioned earlier, NGP interpolation is essential to our implementation of the pp part. At the same time, the biggest problem with the straightforward pp force calculation is that the results are anisotropic and depend on the location of the fine mesh with respect to the particles. As an example, consider two particles on either side of a grid cell boundary, experiencing their mutual gravity attraction via the fine mesh force with a discretized one-grid cell separation. If, instead, the mesh was shifted such that they were within the same cell, they would experience the much larger pp force. This is clearly seen in the top left panel of Fig. 7, where particles physically close, but in different grid cells, feel a force up to an order of magnitude too small. This effect is especially pronounced at the early stages of the simulation where the density is more homogeneous, and leads to mesh artefacts appearing in the density field.

In order to minimize these two systematic effects – the large scatter and the anisotropy – we randomly shift the particle distribution relative to the mesh by a small amount – up to 2 fine grid cells in magnitude – in each dimension and at each time step. This adds negligible computational overhead as it is applied during the particle position update, and suppresses the mesh behaviour that otherwise grows over multiple time steps. It is possible to shift back the particles at the end of each time step, which prevents a random drift of the whole population, a necessary step if one needs to correlate the initial and final positions of the particles for instance, or for hybrid dark matter – MHD simulations.

We ensure that, on average, this solution balances out the mesh feature, by tuning the force kernels such as to provide a force as evenly balanced as possible, both at grid cell distances and at the cutoff length ($r_c = 16$ fine cells). These adjustments are performed from the pairwise force test mentioned above (and described in MPT). The bottom panels of Fig. 7 show the effect of averaging over ten time steps, where the random offset is applied on the particles (and on the hole) at the end of each force calculation. We observe that the agreement extends at the per cent level down to $r \sim 1.5$, beyond which point the discretization effect of the NGP

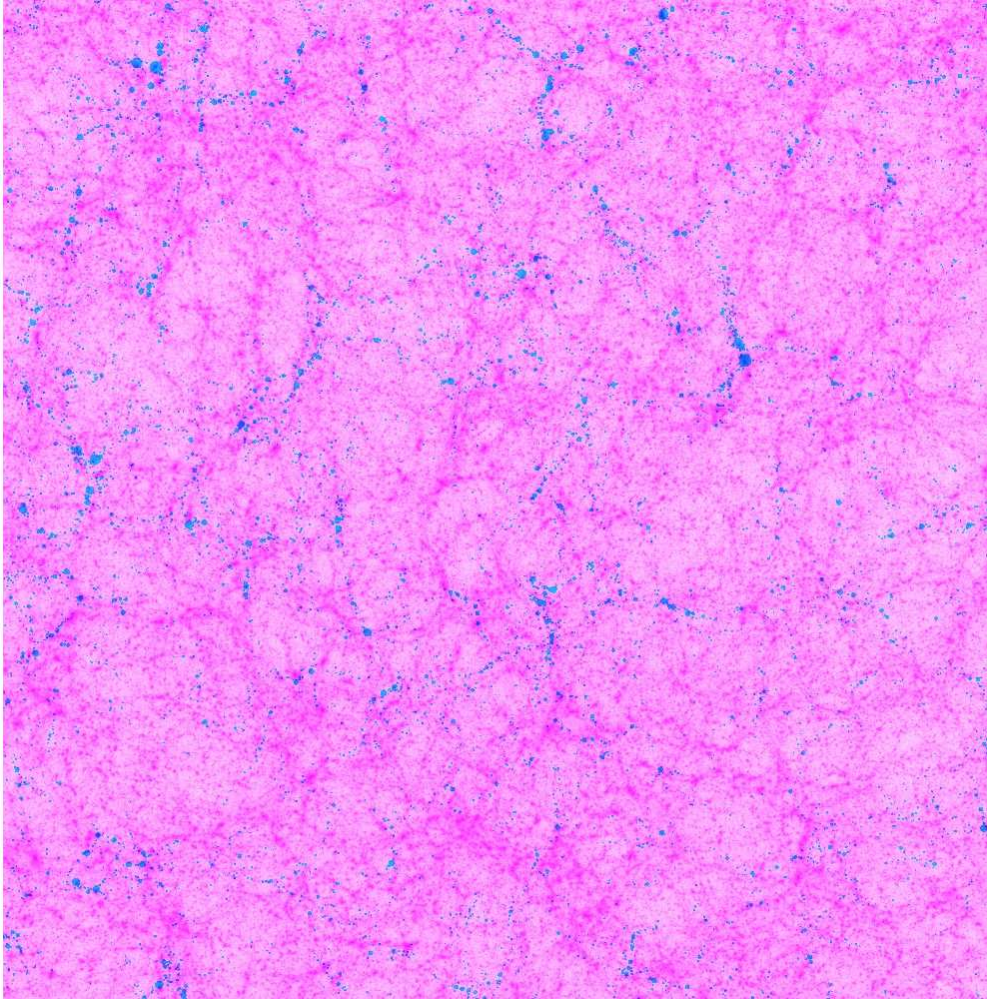


Figure 5. Detail of a dark matter density projection, measured at $z = 0$ in a RANGER4000 simulation. Particles are shown in pink, haloes in blue in the online version. The full projection is 64 times larger, but does not provide much details on the structure on a paper format.

causes a cumulative underestimate of the force, which reaches up 60 per cent.

We note that this inevitable loss of force in the NGP scheme is one of the driving arguments to extend the pp force outside the fine mesh cell, since the scattering about the actual $1/r^2$ law drops rapidly as the distance increases. As discussed in section 7.3, this gain in accuracy comes at a computational price, but at least we have the option to run the code with high precision.

We present in Fig. 8 the dramatic impact of removing the random offset in the otherwise default code configuration. This test was performed with a CITA256 simulation of very large box size, the output redshifts are very early (the upper most curve ($z = 10$) was obtained after about 60 time steps), such that the power spectrum should agree with linear theory up to the resolution limit. Instead, we observe that the power spectrum grows completely wrong, due to the large scatter in the force from the fine mesh, and to the anisotropic nature of the P^3M calculations mentioned above. When these effects are not averaged over, the errors directly add up at each time step, which explains why later times are worst. We recall that PMFAST did not have this problem since it used CIC interpolation on both meshes.

Although most of the code tests presented in this paper were obtained with the configuration described above, we present here

a recent improvement in the fine force kernel that compensates for the systematic underestimate at grid distances. Basically, the idea is to boost the small distance elements of the kernel by a small amount, which somehow corrects for the loss observed in the bottom panels of Fig. 7. We show here the impact of three different trials: a) the force from the first layer of neighbours is boosted by 90 per cent, b) the force from the first two layers of neighbours is boosted by 60 per cent, and c) by 80 per cent. We observe that we are able to improve the power at turn around by almost a factor of two, which doubles the resolution. This test was performed in a series of SciNet256 simulations with a volume of $150h^{-1}\text{Mpc}$ and at $z = 0$. We see from the figure that in the original configuration, the departure from the non-linear predictions occur at $k \sim 3.0h\text{Mpc}^{-1}$, whereas the third trial allows for a resolution down to $k \sim 6.0h\text{Mpc}^{-1}$.

5.3 Constraining redshift jumps

At early stages of the simulation, the density field is homogenous, causing the force of gravity to be rather weak everywhere. In that case, the size of the redshift jumps is controlled by a limit in the cosmological expansion. If the expansion jump is too large, the size of the residual errors can become significant, and one can observe,

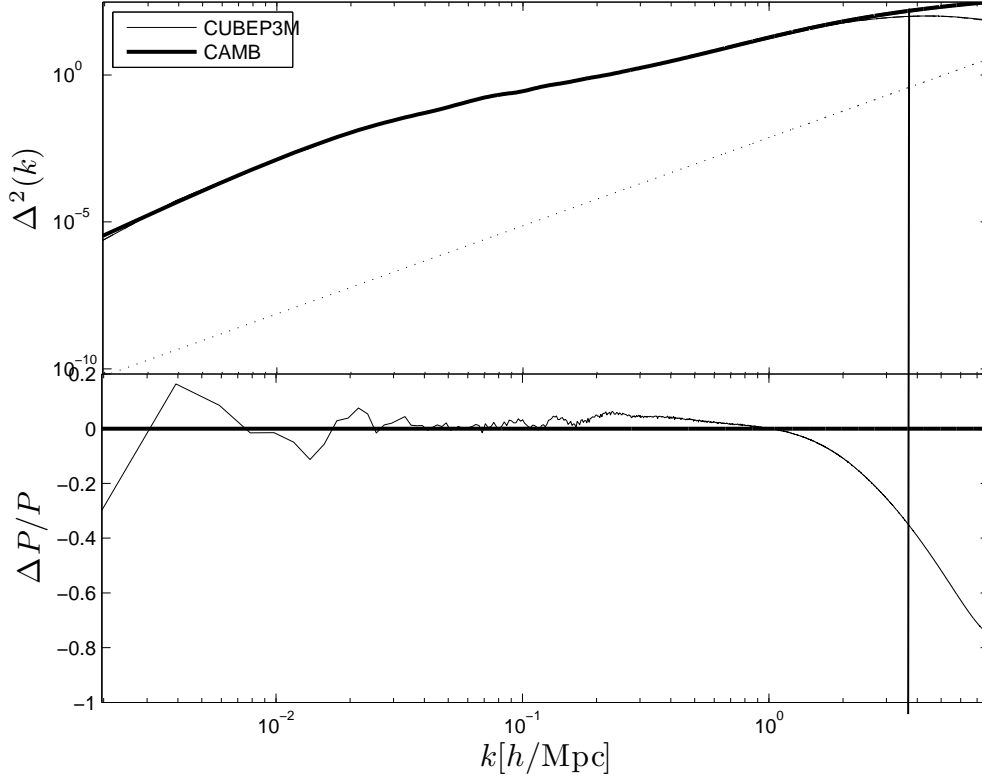


Figure 6. Dark matter power spectrum, measured at $z = 0$ in a RANGER4000 simulation, compared to the non-linear predictions of HALOFIT. The vertical line corresponds to the scale of the coarse mesh, while the dotted line represents the Poisson noise. The particle masses are of $5.68 \times 10^{10} M_{\odot}$.

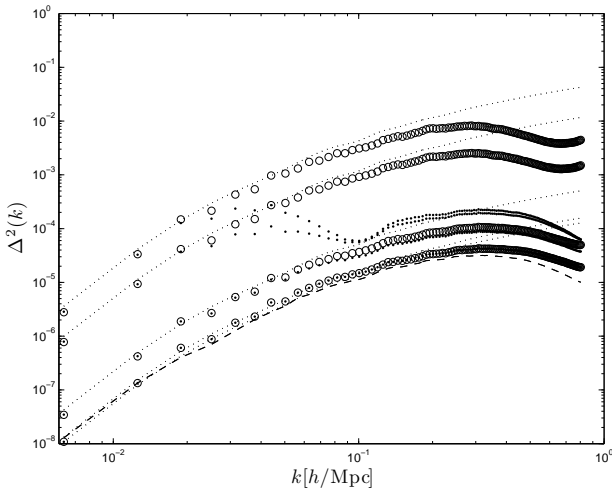


Figure 8. Dark matter power spectrum, measured at $z = 180, 100, 20$ and 10 , in a CITA256 realization that is $1000 \text{ Mpc}/h$ per side. The dashed line represent the initial condition power spectrum, the dotted lines are the linear predictions, and the open circles the standard P^3M configuration. The dots were obtained by simply removing the random offset that is normally applied at each time step.

for instance, a growth of structure that does not match the predictions of linear theory even at the largest scales. One therefore needs to choose a maximum step size. In CUBEP3M, this is controlled by r_{max} , which is the fractional step size, $da/(a + da)$ and is set to 0.05

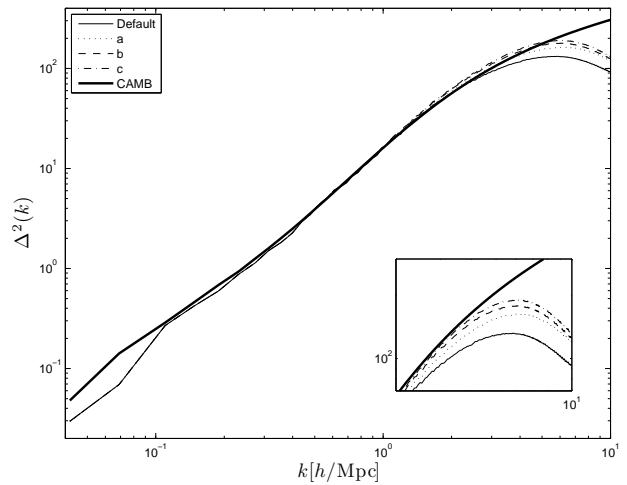


Figure 9. Dark matter power spectrum, measured at $z = 0$, in a SciNet256 realization that is $150 \text{ Mpc}/h$ per side. The thick solid line represents predictions from CAMB, the thin solid line is the default configuration of the code, and the three other lines represent three different corrections that are applied at grid distances to compensate for the systematic force underestimation seen in the bottom panels of Fig. 7.

by default. Generally, a simulation should start at a redshift high enough that the initial dimensionless power spectrum is well under unity at all scales. This ensures that the Zel'dovich approximation holds at the per cent level at least. A drop of accuracy can occur if

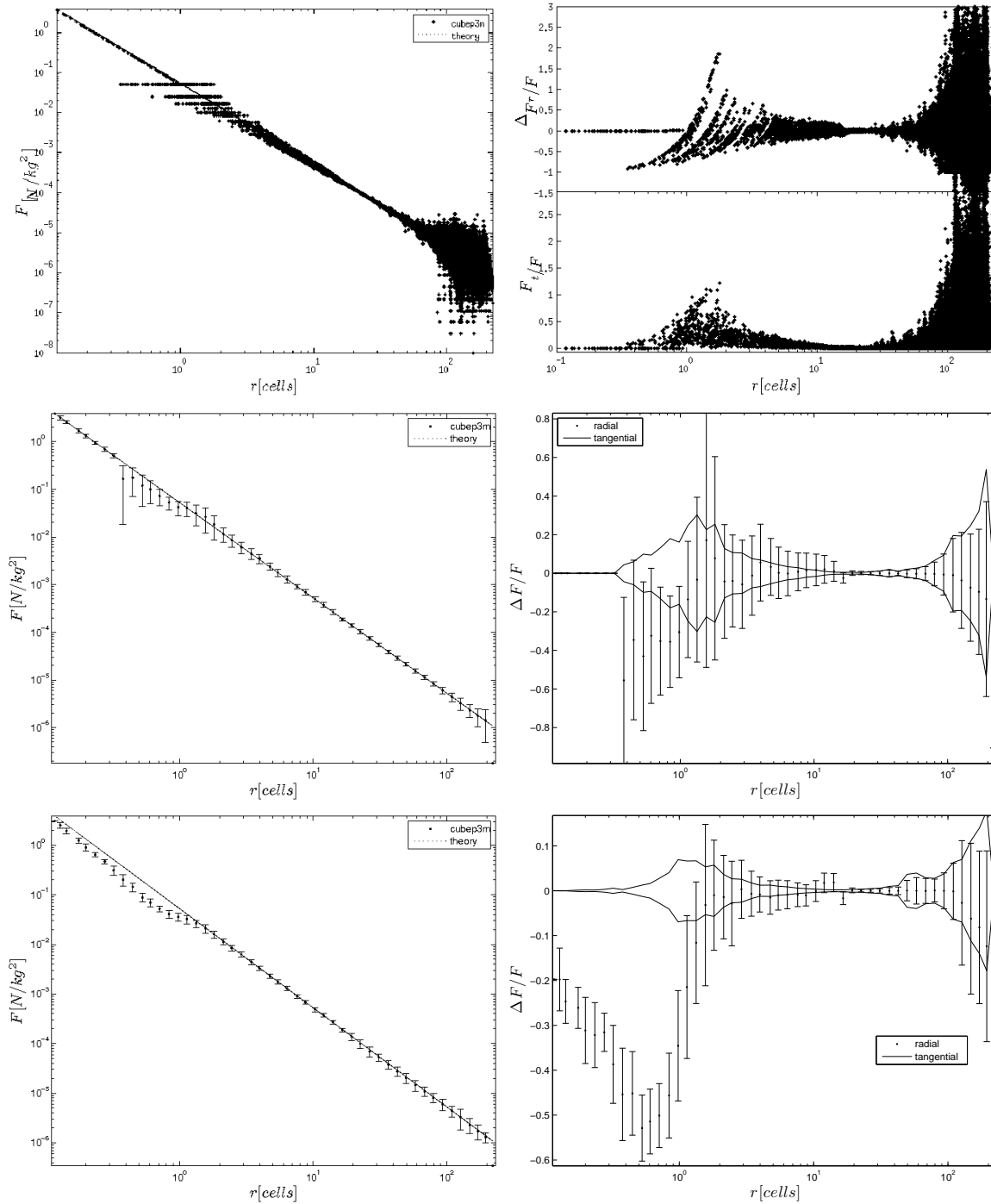


Figure 7. (*top left*.) Gravity force in the P^3M algorithm, versus distance in fine mesh cell units, compared with the exact $1/r^2$ law. This particular calculation was obtained in a CITA128 realization with a box size of $100h^{-1}\text{Mpc}$, in a single time step. (*top right*.) Fractional error on the force in the P^3M algorithm in the radial direction (top) and fractional tangential contribution (bottom). In a full simulation run, the scatter averages over many time steps, thanks to the inclusion of a random offset that is imposed on each particle, as discussed in the main text. (*middle row*.) Top row organized in 50 logarithmic bins; we plot only the scatter about the mean for the tangential contribution. (*bottom row*.) Same as middle row, but averaged over ten time steps.

one starts the simulation too early, where truncation error will be significant at the early time steps.

It is possible to reduce this effect, and thereby improve significantly the accuracy of the code, by modifying the value of r_{max} , at the cost of increasing the total number of time steps. Fig. 10 shows a comparison of late time power spectra of a series of CITA256 realizations that originate from the same initial conditions, and used

the same random seeds to control the fine mesh shifts (mentioned above): only the value of r_{max} was modified between each run. We observe that the impact is mostly located in the non-linear regime, where decreasing the time step to 0.006 allows the simulation to recover about 30 per cent of dimensionless power at the turn over scale, in this simulation configuration. This gain is greatly affected by the choice of initial redshift, the resolution, and the box size,

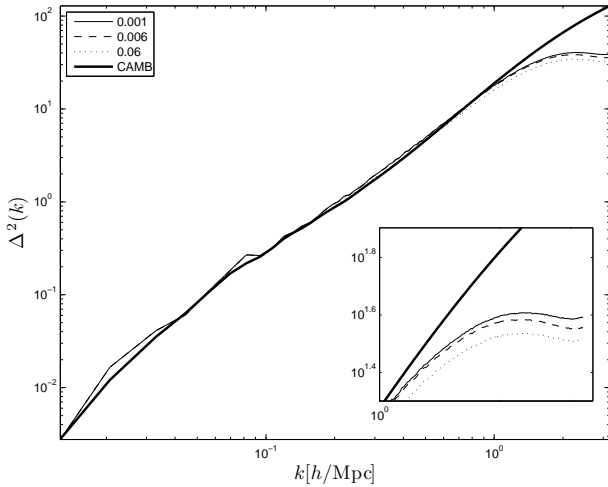


Figure 10. Dark matter power spectrum, measured at $z = 0$ in a series of CITA256 realizations. The starting redshift was raised to $z = 200$ to enhance the systematic effect. The different curves show different values of r_{\max} . The resources required to run these simulations increase rapidly as r_{\max} decreases, as seen in Table 4.

Table 4. Scaling in CPU resources as a function of the value of r_{\max} . The tests were performed on the CITA Sunnyvale cluster, and general trends could vary slightly on other machines. The time tabulated on the right column corresponds to the full run time of the CITA256 simulations, which evolved 128^3 particles from $z = 200$ to $z = 0$.

r_{\max}	time (h)
0.1	1.46
0.06	1.48
0.01	1.67
0.006	1.91
0.003	2.83
0.002	4.13
0.001	8.15

and ideally one would make test runs in order to optimize a given configuration. As expected, the CPU resources required to run these simulations increase rapidly as r_{\max} decreases, as seen in Table 4. In this test case, reducing further at 0.001 shows only a mild improvement in accuracy, but the increase in time is more than a factor of four. The default configuration of the code is set to $r_{\max} = 0.05$, but at the light of this recent test, we recommend reducing to the value 0.01 or even 0.005. We also mention here that with a proper use of second order initial conditions (also see section 7.1), it is possible to start the simulations at much later redshifts without loosing much accuracy.

6 RUNTIME HALO FINDER

We have implemented a halo finding procedure, which we have developed based on the spherical overdensity (SO) approach (Lacey & Cole 1994). In the interest of speed and efficiency the halo catalogues are constructed on-the-fly at a pre-determined list of redshifts. The halo finding is massively-parallel and threaded based on the main CUBEP3M data structures discussed in section 2. The code first builds the fine-mesh density for each sub-domain using CIC or

NGP interpolation. It then proceeds to search for all local density maxima above a certain threshold (typically set to a factor of 100 above mean density) within the local tile. It then uses parabolic interpolation on the density field to determine more precisely the location of the maximum within the densest cell, and records the peak position and value.

Once the list of peak positions is generated, they are sorted from the highest to the lowest density. Then each of the halo candidates is inspected independently, starting with the highest peak. The grid mass is accumulated in spherical shells of fine grid cells surrounding the maximum, until the mean density within the halo drops below a pre-defined overdensity cutoff (usually set to 178 in units of the mean, in accordance with the top-hat collapse model). As we accumulate the mass, we remove it from the mesh, so that no mass element is double-counted. This method is thus inappropriate for finding sub-haloes since within this framework, those are naturally incorporated in their host haloes. Because the haloes are found on a discrete grid, it is possible, especially for those with lower mass, to overshoot the target overdensity. To minimize this effect, we correct the halo mass and radius with an analytical density profile. We use the Truncated Isothermal Sphere (TIS) profile (Shapiro et al. 1999; Iliev & Shapiro 2001) for overdensities below ~ 130 , and a simple $1/r^2$ law for lower overdensities. These yield a similar outer slope to the Navarro, Frenk and White (NFW Navarro et al. 1997) profile, but extends to lower overdensities and matches well the virialization shock position given by the self-similar collapse solution of Bertschinger (1985).

After the correct halo mass, radius and position are determined, we find all particles that are within the halo radius. Their positions and velocities are used to calculate the halo centre-of-mass, bulk velocity, internal velocity dispersion and the three angular momentum components, all of which are then included in the final halo catalogues. We also calculate the total mass of all particles within the halo radius, also listed in the halo data. This mass is very close, but typically slightly lower, than the halo mass calculated based on the gridded density field. The centre-of-mass found this way closely follows that found from the peak location, which is based on the gridded mass distribution.

A sample halo mass function produced from a RANGER4000 simulation at $z = 0$ is shown in Fig. 11. We compare our result to the precise fit presented recently by Tinker et al. (2008). Unlike most other widely-used fits like the one by Sheth & Tormen (2002), which are based on friends-of-friends (FOF) halo finders, this one is based on the SO search algorithm, whose masses are systematically different from the FOF masses (e.g. Reed et al. 2007; Tinker et al. 2008), making this fit a better base for comparison here. Results show excellent agreement, within ~ 10 per cent for all haloes with masses corresponding to 1000 particles or more. Lower-mass haloes are somewhat under-counted compared to the Tinker et al. (2008) fit, by ~ 20 per cent for 400 particles and by ~ 40 per cent for 50 particles. This is largely due to the grid-based nature of our SO halo finder, which misses some of the low-mass haloes. It was shown that using more sophisticated halo finders (available only through post-processing calculations due to their heavier memory footprint) it is possible to recover most of the expected number count.

7 BEYOND THE STANDARD CONFIGURATION

The preceding descriptions and discussions apply to the standard configuration of the code, as described at the beginning of section

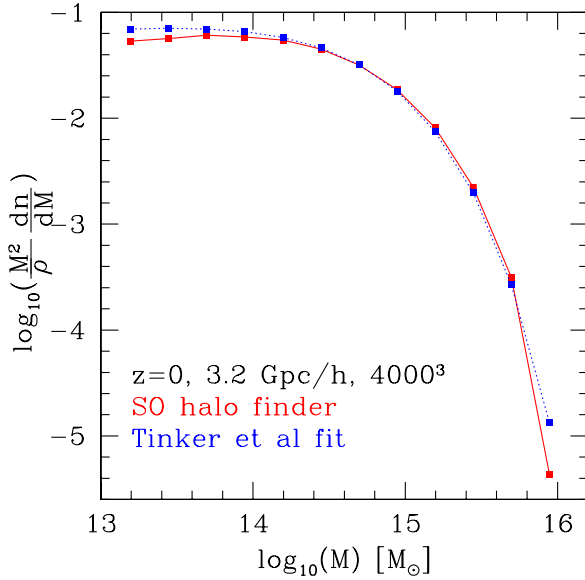


Figure 11. Simulated halo multiplicity function, $\frac{M^2}{\rho} \frac{dn}{dM}$, based on a RANGER4000 simulation with $3.2 h^{-1} \text{Gpc}$ box and 4000^3 particles (solid, red in the online version). For reference we also show a widely-used precise fit by Tinker et al. (2008) (blue, dashed). The particle masses are of $5.68 \times 10^{10} M_{\odot}$.

2. A few extensions have been recently developed in order to enlarge the range of applications of CUBEP3M, and this section briefly describes the most important improvements.

7.1 Initial conditions

As mention in section 2, the code starts off by reading a set of initial conditions. These are organized as a set of $6 \times N$ phase-space arrays – one per MPI process – where N is the number of particles in the local volume.⁷ Although many applications are typically based on initial conditions that follow Gaussian statistics, we have developed a non-Gaussian initial conditions generator that we briefly describe in this section. This becomes important especially for simulations in which the sensitivity to very small scales becomes more important.

The original code that provides Gaussian initial conditions for CUBEP3M is extended to include non-Gaussian features of the ‘local’ form, $\Phi(\mathbf{x}) = \phi(\mathbf{x}) + f_{\text{NL}}\phi(\mathbf{x})^2 + g_{\text{NL}}\phi(\mathbf{x})^3$, where $\phi(\mathbf{x})$ is the Gaussian contribution to the Bardeen potential $\Phi(\mathbf{x})$ (see Bartolo et al. (2004) for a review). We adopted the CMB convention, in which Φ is calculated immediately after the matter-radiation equality (and not at redshift $z = 0$ as in the large scale structure convention). For consistency, $\phi(\mathbf{x})$ is normalized to the amplitude of scalar perturbations inferred by CMB measurements ($A_s \approx 2.2 \times 10^{-9}$). The local transformation is performed before the inclusion of the

⁷ We remind that for Zel’dovich approximation to hold in such cases, the simulations need to be started at very early redshifts. Consequently, the size of the first few redshift jumps in such simulations can become rather large, and therefore less accurate, which is why we must carefully constrain the size of the jumps, as discussed in section 5.3. Another solution to this issue is to use second order perturbation theory to generate the initial conditions, which is not implemented yet in the public package.

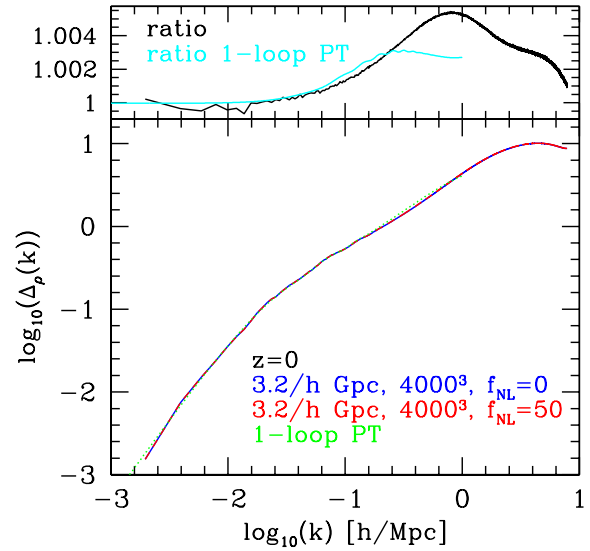


Figure 12. Dark matter power spectrum, measured at $z = 0$ in a volume $3.2 h^{-1} \text{Gpc}$ per side, from 4000^3 particles. The two curves represent two RANGER4000 realizations of the same initial power spectrum, one of which used Gaussian statistics (blue) and the other the non-Gaussian initial condition generator. The two curves differ at the sub-per cent level, as seen in the top panel, and the one-loop perturbation theory calculations accurately describes the ratio between the two curves up to $k \sim 0.4 h \text{Mpc}^{-1}$ in this case.

matter transfer function, and the initial particle positions and velocities are finally computed from $\Phi(\mathbf{x})$ according to the Zel’dovich approximation, as in the original Gaussian initial condition generator.

This code was tested by comparing simulations and theoretical predictions for the effect of local primordial non-Gaussianity on the halo mass function and matter power spectrum (Desjacques, Seljak & Iliev 2009). It has also been used to quantify the impact of local non-Gaussian initial conditions on the halo power spectrum (Desjacques et al. 2009; Desjacques & Seljak 2010) and bispectrum (Sefusatti et al. 2010), as well as the matter bispectrum (Sefusatti et al. 2011). Fig. 12 shows the late time power spectrum of two RANGER4000 realizations that started off the same initial power spectrum, but one of which had non-Gaussian initial conditions. We see that the difference between the two power spectra is at the sub-per cent level, and that the ratio of the two power spectra are well described with one loop perturbation theory (Scoccimarro et al. 2004; Taruya et al. 2008).

7.2 Particle identification tags

A system of particle identification can be turned on, which basically allows to track each particle’s trajectory between checkpoints. Such a tool is useful for a number of applications, from reconstruction of halo merging history to tracking individual particle trajectories. The particle tag system has been implemented as an array of double integers, PID, and assigns a unique integer to each particle during the initialization stage. The location of the tag on the PID array matches the location of the corresponding particle on the xv array, hence it acts as if the latter array had an extra dimension. The location change only when particles exist the local volume, in which

case the tag is sent along adjacent nodes with the particle in the `pass_particle` subroutine, and deleted particles result in deleted flags. Similarly to the phase space array, the PID array gets written to file at each particle checkpoint. Since the array takes a significant amount of memory, we opted to store the tags in a separate object – as opposed to adding an extra dimension to the existing `xv` array, such that it can be turned off when memory becomes an issue.

7.3 Extended range of the pp force

One of the main sources of error in the calculation of the force occurs from the PM interaction at the smallest scales of the fine grid. The approximation by which particles in a neighbouring mesh grid can be placed at the centre of the cell is less accurate, which causes a maximal scatter around the exact $1/r^2$ law. A solution to minimize this error consists in extending the pp force calculation outside a single cell, which inevitably reintroduces a number of operations that scales as N^2 . Our goal is to add the flexibility to have a code that runs slower, but produces results with a higher precision.

To allow this feature, we have to choose how far outside a cell we want the exact pp force. Since the force kernels on both meshes are organized in terms of grids, the simplest way to implement this feature is to shut down the mesh kernels in a region of specified size, and allow the pp force to extend therein. Concretely, these regions are constructed as cubic layers of fine mesh grids around a central cell; the freedom we have is to choose the number of such layers.

To speed up the access to all particles within the domain of computation, we construct a thread safe linked list to be constructed and accessed in parallel by each core of the system, but this time with a head-of-chain that points to the first particle in the current fine mesh cell. We then loop over all fine grids, accessing the particles contained therein and inside each fine grid cell for which we killed the mesh kernels, we compute the separation and the force between each pair and update their velocities simultaneously with Newton's third law. To avoid double counting, we loop only over the fine mesh neighbours that produce non-redundant contributions. Namely, for a central cell located at (x_1, y_1, z_1) , we only consider the neighbours (x_2, y_2, z_2) that satisfy the following conditions:

- $z_2 \geq z_1$ always
- if $z_2 = z_1$, then $y_2 \geq y_1$, otherwise we also allow $y_2 < y_1$
- if $z_2 = z_1$ and $y_2 = y_1$, then we enforce $x_2 > x_1$

The case where all three coordinates are equal is already calculated in the standard configuration of the code. To assess the improvement of the force calculation, we present in Fig 13 a force versus distance plot in a CITA128 realization, analogous to Fig. 7, but the pp force has been extended to two layers of fine cells in the force test part of the code. We observe that the scatter about the theoretical curve has reduced significantly, down to the few percent level, and is still well balanced around the theoretical predictions. The fractional error on the radial and tangential components of the force, as seen in the right panel, are now at least five times smaller than in the default P^3M algorithm. When averaging over 10 time steps, we observe that improvement is relatively mild, as the calculation is already very accurate.

To quantify the accuracy improvement versus computing time requirements, we performed the following test. We generate a set of initial conditions at a starting redshift of $z = 50$, with a box size equal to $150h^{-1}\text{Mpc}$, and with 512^3 particles. We evolve these

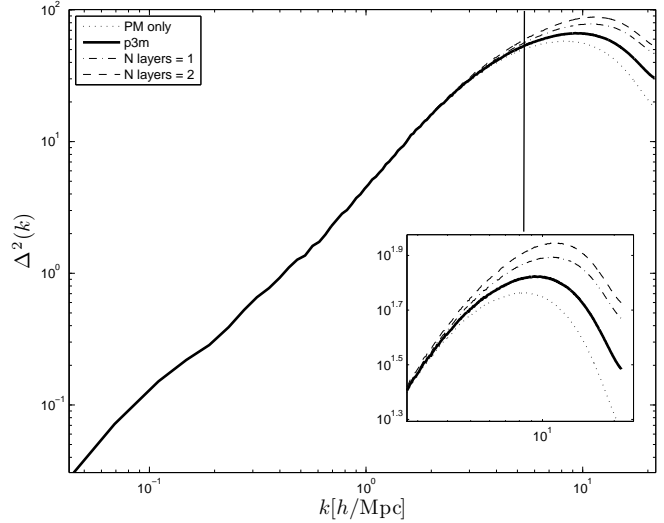


Figure 14. Dimensionless power spectrum for varying range of the exact pp force. These SciNet512 realizations evolved from a unique set of initial conditions at a starting redshift of $z = 100$, with a box size equal to $150h^{-1}\text{Mpc}$ until $z = 1.0$. The only difference between the runs is the ranges of the pp calculation. The inset shows details about the resolution turnaround, and the thick vertical line corresponds to the coarse mesh scale.

SciNet512 realizations with different ranges for the pp calculation, and compare the resulting power spectra. For the results to be meaningful, we also need to use the same random seed for the random number generator, such that the only difference between different runs is the range of the pp force. Fig. 14 shows the dimensionless power spectrum of the different runs, where we see a significant gains in resolution when extending PM to P^3M first, and when adding successively one and two layers of fine cell mesh in which the pp force is extended. We have not plotted the results for higher numbers of layers, as the improvement becomes milder there: the mesh calculations become accurate enough as the distance increases. For this reason, it seems that a range of two layers suffices to reduce most of the undesired NGP scatter.

Extending the pp calculation comes at a price, since the number of operations scales as N^2 in the sub-domain. This cost is best captured by the increase of real time required by a fixed number of dedicated CPUs to evolved the particles to the final redshift. For instance, in our SciNet512 simulations, the difference between the default configuration and $N_{\text{layer}} = 1$ is about a factor of 2.78 in real time, and about 6.5 for $N_{\text{layer}} = 2$. This number will change depending on the problem at hand and on the machine, and we recommend to perform performance gain vs resource usage tests on smaller runs before performing large scale simulations with the extended force.

The power spectrum does not provide the complete story, and one of the most relevant ways to quantify the improvement of the calculation is to measure the impact on the halo mass function from these different SciNet512 runs. Fig. 15 presents this comparison at redshift $z = 1.0$. About 76,000 haloes were found in the end, yielding a halo number density of about 0.0225 per Mpc/h^3 . We observe that the simulation undershoot the Sheth-Tormen predictions, which is caused by the relatively low resolution of the configuration compared to the RANGER4000 run. We also observe that the pure PM code yields up to 10 per cent less haloes than the P^3M

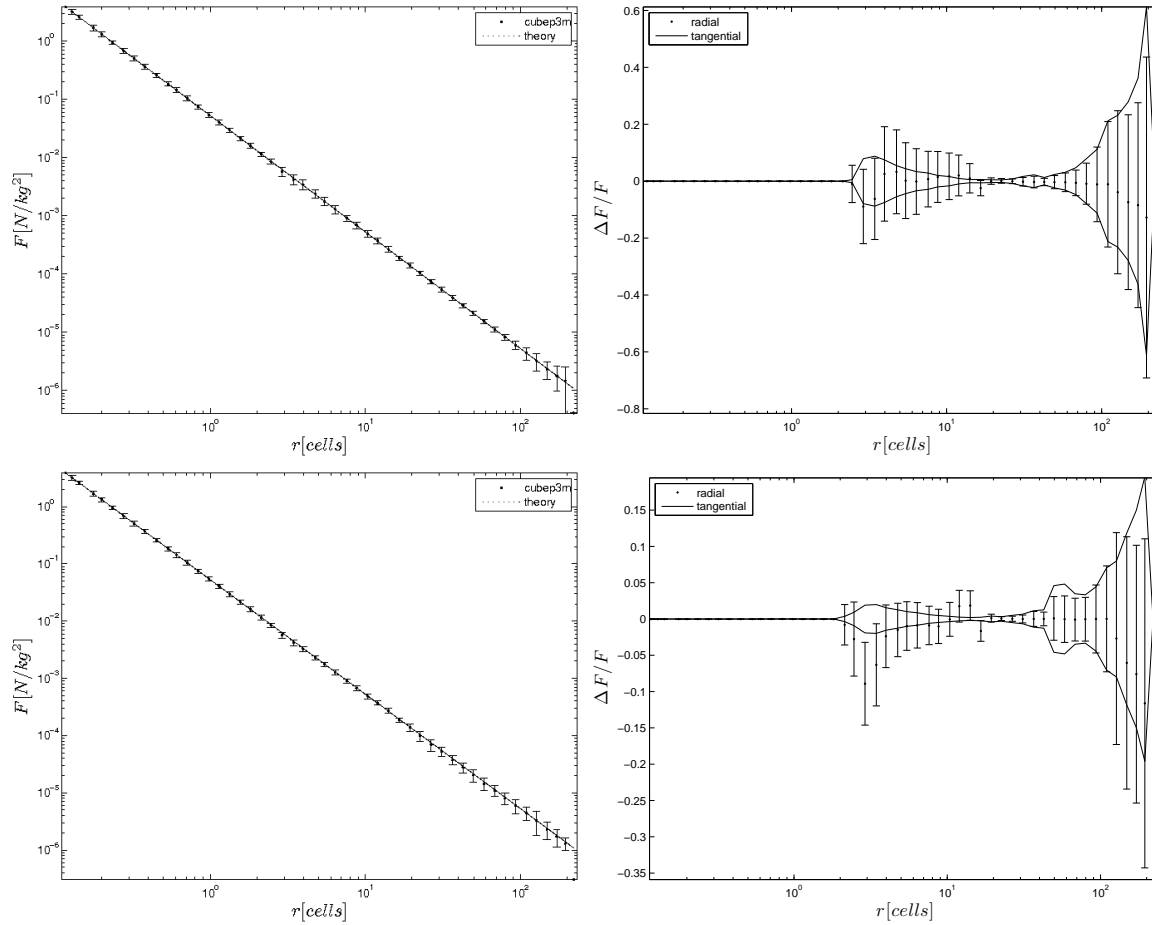


Figure 13. (top left:) Gravity force in the P³M algorithm, compared with the exact $1/r^2$ law, in the same CITA128 realization as that shown in Fig. 7, except that the exact pp force has been extended to two fine mesh layers around each particle in the force test code. Particles in that range follow the exact curve, then we observe a much smaller scatter at distance of the order of 2 fine grid, compared to Fig. 7. The NGP interpolation scheme is again responsible for the scatter, but the effect is suppressed with increasing distances, and further suppressed by averaging over many time steps. (top right:) Fractional error on the force in the P³M algorithm, in the radial direction (top) and fractional tangential contribution (bottom). This was also obtained over a single time step. (bottom row:) Same as top row, but averaged over ten time steps.

version, over most of the mass range, whereas the extended pp algorithm generally recovers up to 10 per cent more haloes in the range $10^{11} - 10^{13}$. The difference in performance between an extended pp range of two vs four cells is rather mild, from where we conclude that $N_{\text{layers}} = 2$ is the most optimal choice.

8 CONCLUSION

This paper describes CUBEP3M, a public and massively parallel P³M N-body code that inherits from PMFAST and that now scales well to 20,000 cores, pushing the limits of the cosmological problem size one can handle. We summarize the code structure, review the double-mesh Poisson solver algorithm, and present scaling and systematic tests that have been performed. We also describe various utilities and extensions that come with the public release, including a run time halo finder, an extended pp force calculation and a non-Gaussian initial condition generator. CUBEP3M is one of the most competitive N-body code that is publicly available for cosmologists and astrophysicists, it has already been used for a large number of scientific applications, and it is our hope that the current documentation will help the community in interpreting its outcome. The

code is publicly available on github.com under `cubep3m`, and extra documentation about the structure, compiling and running strategy is can be found on the CITA wiki page⁸.

ACKNOWLEDGEMENTS

The CITA simulations were run on the Sunnyvale cluster at CITA. ITI was supported by The Southeast Physics Network (SEPNet) and the Science and Technology Facilities Council grants ST/F002858/1 and ST/I000976/1. Computations for the SciNet1024 runs were performed on the GPC supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto. The authors acknowledge the TeraGrid and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin (URL: <http://www.tacc.utexas.edu>) for providing HPC and visualization

⁸ wiki.cita.utoronto.ca/mediawiki/index.php/CubePM

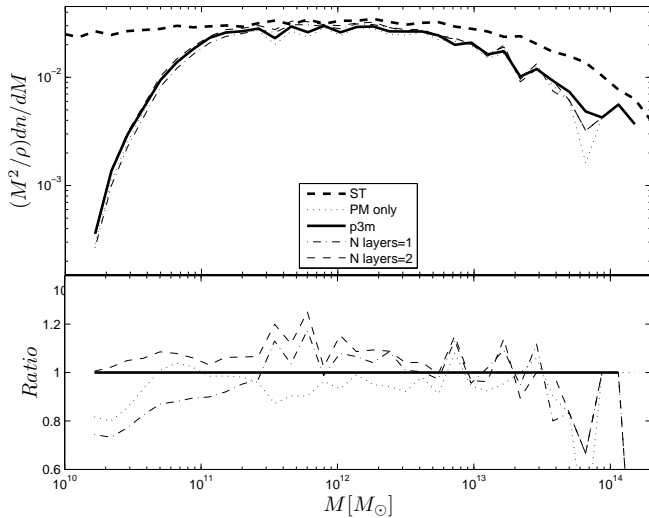


Figure 15. (*top:*) Halo mass function for different ranges of the pp force calculation, compared to the predictions of Sheth-Tormen, at $z = 1.0$. The smallest haloes shown here have a mass equivalent to 100 particles, and fall in the lowest mass bin. (*bottom:*) Ratio between the different curves and that of the default P³M configuration.

resources that have contributed to the research results reported within this paper. ITI also acknowledges the Partnership for Advanced Computing in Europe (PRACE) grant 2010PA0442 which supported the code scaling studies. UEP and JDE are supported by the NSERC of Canada. VD acknowledges support by the Swiss National Science Foundation.

REFERENCES

- Bartolo N., Komatsu E., Matarrese S., Riotto A., 2004, *Phys. Rep.*, 402, 103
- Bertschinger E., 1985, *ApJS*, 58, 39
- Colless M., et al., 2003, ArXiv e-prints (astro-ph/0306581)
- Couchman H. M. P., 1991, *ApJ*, 368, L23
- Couchman H. M. P., Thomas P. A., Pearce F. R., 1995, *ApJ*, 452, 797
- Datta K. K., Mellema G., Mao Y., Iliev I. T., Shapiro P. R., et al., 2011, arXiv:1109.1284
- Desjacques V., Seljak U., 2010, *Phys. Rev. D*, 81, 023006
- Desjacques V., Seljak U., Iliev I. T., 2009, *MNRAS*, 396, 85
- Doré O., Lu T., Pen U.-L., 2009, ArXiv e-prints
- Drinkwater M. J., et al., 2010, *MNRAS*, 401, 1429
- Dubinski J., Kim J., Park C., Humble R., 2004, *NewA*, 9, 111
- Dyer C. C., Ip P. S. S., 1993, *ApJ*, 409, 60
- Eisenstein D. J., et al., 2005, *ApJ*, 633, 560
- Eleftheriou M., Fitch B. G., Rayshubskiy A., Ward T. J. C., Germain R. S., 2005, *IBM J. Res. Dev.*, 49, 457
- Fernandez E. R., Iliev I. T., Komatsu E., Shapiro P. R., 2011, arXiv:1112.2064
- Friedrich M. M., Datta K. K., Mellema G., Iliev I. T., 2012, ArXiv e-prints
- Friedrich M. M., Mellema G., Alvarez M. A., Shapiro P. R., Iliev I. T., 2011, *MNRAS*, 413, 1353
- Frigo M., Johnson S., 2005, in *Proceedings of the IEEE Vol. 93, The Design and Implementation of FFTW3*. pp 216–231

- Harnois-Déraps J., Pen U.-L., 2012, *MNRAS*, 423, 2288
- Harnois-Déraps J., Vafaei S., Van Waerbeke L., 2012, ArXiv e-prints
- Harnois-Déraps J., Yu H.-R., Zhang T.-J., Pen U.-L., 2012, ArXiv e-prints
- Heymans C., CFHTLenS Collaboration 2012, in *American Astronomical Society Meeting Abstracts Vol. 219 of American Astronomical Society Meeting Abstracts, The CFHT Lensing Survey*. p. 130.01
- Hilbert S., Hartlap J., White S. D. M., Schneider P., 2009, *A&A*, 499, 31
- Hockney R. W., Eastwood J. W., 1981, *Computer Simulation Using Particles*
- Ilie C., Freese K., Valluri M., Iliev I. T., Shapiro P. R., 2012, *MNRAS*, p. 2794
- Iliev I. T., Ahn K., Koda J., Shapiro P. R., Pen U.-L., 2010, proceedings paper for Moriond 2010 meeting (ArXiv:1005.2502)
- Iliev I. T., Mellema G., Pen U.-L., Merz H., Shapiro P. R., Alvarez M. A., 2006, *MNRAS*, 369, 1625
- Iliev I. T., Mellema G., Shapiro P. R., Pen U.-L., Mao Y., Koda J., Ahn K., 2011, ArXiv e-prints
- Iliev I. T., Shapiro P. R., 2001, *MNRAS*, 325, 468
- Iliev I. T., Shapiro P. R., Mellema G., Merz H., Pen U.-L., 2008, in refereed proceedings of TeraGrid08, ArXiv e-prints (0806.2887)
- Komatsu E., Dunkley J., Nolte M. R., Bennett C. L., Gold B., Hinshaw G., Jarosik N., Larson D., Limon M., Page L., Spergel D. N., Halpern M., Hill R. S., Kogut A., Meyer S. S., Tucker G. S., Weiland J. L., Wollack E., Wright E. L., 2009, *ApJS*, 180, 330
- Komatsu E., et al., 2011, *ApJS*, 192, 18
- Lacey C., Cole S., 1994, *MNRAS*, 271, 676
- Loken C., et al., 2010, *Journal of Physics: Conference Series*, 256
- Lu T., Pen U.-L., 2008, *MNRAS*, 388, 1819
- Lu T., Pen U.-L., Doré O., 2010, *Phys. Rev. D*, 81, 123015
- Mao Y., Shapiro P. R., Mellema G., Iliev I. T., Koda J., Ahn K., 2012, *MNRAS*, 422, 926
- Mellema G., Iliev I. T., Alvarez M. A., Shapiro P. R., 2006, *New Astronomy*, 11, 374
- Merz H., Pen U.-L., Trac H., 2005, *New Astronomy*, 10, 393
- Navarro J. F., Frenk C. S., White S. D. M., 1997, *ApJ*, 490, 493
- Ngan W., Harnois-Déraps J., Pen U.-L., McDonald P., MacDonald I., 2012, *MNRAS*, 419, 2949
- Pang B., Pen U.-L., Perrone M., 2010, ArXiv e-prints
- Pen U.-L., Arras P., Wong S., 2003, *ApJS*, 149, 447
- Percival W. J., Cole S., Eisenstein D. J., Nichol R. C., Peacock J. A., Pope A. C., Szalay A. S., 2007, *MNRAS*, 381, 1053
- Reed D. S., Bower R., Frenk C. S., Jenkins A., Theuns T., 2007, *MNRAS*, 374, 2
- Rimes C. D., Hamilton A. J. S., 2005, *MNRAS*, 360, L82
- Sato M., Hamana T., Takahashi R., Takada M., Yoshida N., Matsumura T., Sugiyama N., 2009, *ApJ*, 701, 945
- Schlegel D., White M., Eisenstein D., 2009, ArXiv e-prints (astro-ph/0902.4680)
- Scoccimarro R., Sefusatti E., Zaldarriaga M., 2004, *Phys. Rev. D*, 69, 103513
- Sefusatti E., Crocce M., Desjacques V., 2010, *MNRAS*, 406, 1014
- Sefusatti E., Crocce M., Desjacques V., 2011, ArXiv e-prints
- Shapiro P. R., Iliev I. T., Raga A. C., 1999, *MNRAS*, 307, 203
- Sheldon E. S., Johnston D. E., Masjedi M., McKay T. A., Blanton M. R., Scranton R., Wechsler R. H., Koester B. P., Hansen S. M., Frieman J. A., Annis J., 2009, *ApJ*, 703, 2232
- Sheth R. K., Tormen G., 2002, *MNRAS*, 329, 61

- Smith R. E., Peacock J. A., Jenkins A., White S. D. M., Frenk C. S., Pearce F. R., Thomas P. A., Efstathiou G., Couchman H. M. P., 2003, *MNRAS*, 341, 1311
- Springel V., 2005, *MNRAS*, 364, 1105
- Springel V., Yoshida N., White S. D. M., 2001, *New Astronomy*, 6, 79
- Takahashi R., et al., 2009, *ApJ*, 700, 479
- Takahashi R., et al., 2011, *ApJ*, 726, 7
- Taruya A., Koyama K., Matsubara T., 2008, *Phys. Rev. D*, 78, 123534
- Tegmark M., et al., 2006, *Phys. Rev. D*, 74, 123507
- Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, *ApJ*, 688, 709
- Trac H., Cen R., 2007, *ApJ*, 671, 1
- Trac H., Pen U., 2003, in *American Astronomical Society Meeting Abstracts Vol. 35 of Bulletin of the American Astronomical Society, Out-of-Core Hydrodynamic Simulations of the IGM*. p. 1365
- Trac H., Pen U.-L., 2004, *NewA*, 9, 443
- Vafaei S., Lu T., van Waerbeke L., Semboloni E., Heymans C., Pen U.-L., 2010, *Astroparticle Physics*, 32, 340
- Vale C., White M., 2003, *ApJ*, 592, 699
- Xu G., 1995, *ApJS*, 98, 355
- York D. G., et al., 2000, *AJ*, 120, 1579
- Yu H., Zhang T., Harnois-Déraps J., Pen U., 2010, *ArXiv e-prints*
- Zackrisson E., Scott P., Rydberg C.-E., Iocco F., Sivertsson S., Östlin G., Mellema G., Iliev I. T., Shapiro P. R., 2010, *MNRAS*, 407, L74
- Zahn O., Lidz A., McQuinn M., Dutta S., Hernquist L., Zaldarriaga M., Furlanetto S. R., 2007, *ApJ*, 654, 12
- Zhang T., Yu H., Harnois-Déraps J., MacDonald I., Pen U., 2010, *ArXiv e-prints*

This paper has been typeset from a \LaTeX file prepared by the author.